The average height of binary trees and other simple trees

Philippe Flajolet, Andrew M. Odlyzko

INRIA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. 954 90 20

Rapports de Recherche

N° 56

THE AVERAGE HEIGHT OF BINARY TREES AND OTHER SIMPLE TREES

Philippe FLAJOLET
Andrew ODLYZKO

Février 1981

THE AVERAGE HEIGHT OF BINARY TREES AND OTHER SIMPLE TREES[†]

Philippe FLAJOLET
INRIA
78150 Rocquencourt
FRANCE

Andrew ODLYZKO
Bell Laboratories
Murray Hill, New Jersey 07974
U.S.A.

[†] A preliminary version of this work has been presented at the 21st Annual Symposium on Foundations of Computer Science, Syracuse, October 13-15, 1980.

Résumé :

La hauteur moyenne d'un arbre binaire vaut aymptotiquement $2\sqrt{\pi n}$. Cette quantité représente également la taille moyenne de la pile de récursion de l'algorithme de parcours récursif d'arbres. La méthode utilisée s'applique au parcours d'arbres unaires-binaires, d'arbres 2-3 non-équilibrés, d'arbres t-aires et de nombreuses familles simples d'arbres. On obtient comme cas particulier les deux résultats précédemment connus de hauteur moyenne d'arbres : pour les arbres étiquetés, le résultat de Renyi et Szekeres ; pour les arbres planaires, le résultat de De Bruijn, Knuth et Rice. La méthode développée ici qui repose sur une analyse de singularités des séries génératrices est de portée générale.

Abstract :

The average height of a binary tree with n internal nodes is shown to be asymptotic to $2\sqrt{\pi n}$. This represents the average stack height of the simplest recusive tree traversal algorithm. The method used is also applicable to the analysis of traversal algorithms of unary-binary trees, unbalanced 2-3 trees, t-ary trees for any t and other families of trees. It yields the two previously known estimates about average height of trees, namely for labelled non-planar trees a result due to Renyi and Szekeres and for planar trees a resut of De Bruijn, Knuth and Rice. The method developped here, which relies on a singularity analysis of intervening generating fuctions,is new and of a wide applicability

INTRODUCTION

We consider here the problem of the relation between height and size in trees, for various types of trees.

Given a family F of trees with F_n the subset of those trees formed with n nodes, the problem is to determine the average heights defined by

$$H_n(F) = \frac{1}{\text{card } F_n} \sum_{t \in F_n} \text{height}(t).$$

This paper solves the problem for the family B of binary trees, and we show :

THEOREM B : *The average height of binary trees with* n *internal nodes is asymptotically*

$$H_n(B) \sim 2\sqrt{\pi n} \quad \text{as} \quad n \to \infty .$$

So far the only result avaibable about average height of planar trees deals with the family G of general trees ; i.e. planar trees with unrestricted node degrees [2] :

THEOREM G : [De Bruijn, Knuth, Rice] *The average height of general planar trees (of arbitrary node specification) with* n *nodes satisfies*

$$H_n(G) \sim \sqrt{\pi n} \quad \text{as} \quad n \to \infty .$$

The similarities in the forms of Theorem P and Theorem B might induce the reader to believe that Theorem B is only a simple modification of Theorem P. However the methods differ in an essential way.

Theorem P is proved by first giving exact enumerations for the number of trees of fixed height and fixed size ; these are expressed as certain sums of binomial coefficients. The asymptotics are then performed by appealing to properties of the Mellin integral transform and this method is an important starting point of a number of analyses [9] amongst which we mention : radix exchange sort, digital search, Patricia trees, sorting networks and register allocation.

The problem we encounter with binary trees is rather different : the difficulty lies in that exact enumeration formulae are no longer available for the number of trees of fixed size and height. The path we follow is quite general : it relies on the principle that the coefficients of a generating function are largely determined by the location and nature of its singularities. It is also the only recourse when one has at one's disposal nothing but functional equations over generating functions.

The power of the method is due to the fact that most enumeration problems can be translated into functional equations of some sort over generating functions. Locating singularities is achieved by applying approximations and obtaining asymptotic expansions in the complex plane. Coefficients of generating functions are then recovered from these expressions by means of contour integration.

Despite its power this method has only been scarcely used in algorithmic analyses. The work closest to ours is certainly the determination by Odlyzko of the number of balanced 2-3 trees [12]. We demonstrate the generality of our approach by showing :

THEOREM S : *For each simple family of trees S there exists an effectively computable constant* $c(S)$ *such that the average height of a tree in S with* n *nodes is*

$$\overline{H}_n(S) \sim c(S)\sqrt{\pi n}.$$

A family of trees is said to be simple if, essentially, for each r there is a finite set of allowable labels for nodes of degree r. Theorem S is of a very general applicability. It contains as subcases the result by De Bruijn, Knuth and Rice on the average height of planar trees, and -though it does not immediately fits into our framework- a result by Renyi and Szekeres about non planar labelled trees.

Since the height of a tree represents the stack size needed in recursively traversing the tree, Theorem S also yields the analysis of the simplest recursive tree traversal algorithm in a diversity of contexts. The reader should however be warned that statistics on binary search trees represent a different problem to be briefly discussed later.

To conclude this introduction, we should like to emphasize that the interest of the paper is also, as we feel, largely methodological. Almost all "classical" analyses of algorithms follow a chain starting with exact enumeration formulae derived by "direct" counting arguments continued by real approximations (usually approximating discrete sums by integrals). There is a very clear stage at which this approach fails to apply : either the nature of the problem leads to combinatorial expression whose estimation proves intractable, or even more plainly -as is the case here- no combinatorial expression is available at all. In both cases, studying the analytical properties of the corresponding generating function -especially their singularities- leads to solution of problems intrisically not tractable by more elementary methods.

The plan of the paper is as follows :

In the binary case, a certain generating function of the H_n, $H(z)$, is shown to be the sum of quantities defined by a quadratic recurrence (section 2). Recovering the H_n from $H(z)$ requires a detailed analytical investigation of the behaviour of $H(z)$ and in particular necessitates continuing $H(z)$ outside its circle of convergence (section 3) and studying the nature of $H(z)$ around its singularity at $z=1/4$ (section 4).

It is then shown that in an appropriate neighbourhood of $1/4$, $H(z)$ is the sum of a logarithmic term and of a remainder term with smaller order. Most of the difficulty of the problem comes from deriving this expansion.

The coefficients H_n are then obtained from $H(z)$ by means of the Cauchy residue theorem (section 5) with the choice of an adequate contour of integration similar to the one used in the study of balanced 2-3 trees by Odlyzko [12]. The contour is taken to give predominance to the behaviour of the function around its singularity.

We indicate how to extend the method to any simple family of trees (section 6). This includes all previously known results about the height of trees and provides the very general result stated in Theorem S. Last (section 7) we discuss the limits of the present approach and some of its extensions to estimates of higher moments and limit distributions.

I - TREE TRAVERSAL

We shall limit ourselves here to a short algorithmic discussion of tree traversal, refering the reader to [8] for more details.

Perhaps one of the simplest recursive algorithms is the algorithm for visiting -one also says traversing or exploring- nodes of a planar tree. The algorithm occurs in a number of contexts in compiling, program transformation, term rewriting systems, optimization... . Loosely described, this simple algorithm looks like

```
procedure VISIT(T:tree)
      do-something-with(root(T)) ;
      for U subtree-of-root-of T do
          VISIT (U)
      rof
erudecorp
```

In specific applications, the trees input to the algorithm usually obey some particular format. For instance one may encounter : expression trees involving nullary symbols (variables), unary symbols (log,sin,$\sqrt{}$) and binary symbols ($+,-,\times,\div$) ; syntax trees of various types with nodes of possibly unbounded degrees (as in list-of-instruction nodes) ; trees to represent terms in formal manipulation systems... .

We are interested here in the behaviour of the tree exploration procedure in such various contexts.

The analysis of the time of the VISIT procedure is not difficult since the complexity is clearly linear in the size of the input tree. The main problem is to evaluate storage utilization, i-e to determine the average stack size -equivalently recursion depth- required for exploring a tree, as a function of the size of the tree. For a given tree, the stack size required by the visit is equal to the height of the tree. Average case analysis of the algorithm applied to a family F of input trees thus reduces to determining average heights of trees in F.

The available results about average heights of planar trees, old and new, have been described in the introduction. Results of this paper thus completely solve the average case analysis of tree traversal applied to any simple family of inputs. In particular, Theorem B can be rephrased as :

THEOREM B' : *The recursive traversal procedure applied to binary trees of size* n, *has average storage complexity*

$$H_n(\mathcal{B}) \sim 2\sqrt{\pi n} \qquad as \quad n \to \infty .$$

It is to be mentionned here that the result by De Bruijn, Knuth and Rice relative to the family $\mathcal{S}$ of general planar trees, namely that

$$H_n(\mathcal{G}) = \sqrt{\pi n} - \frac{1}{2} + o(1)$$

gives some information on the height of binary trees, as well as on binary tree traversal.

Indeed the rotation correspondence ([8], section 2.3.2) transforms a general tree with n nodes into a binary tree containing (n-1) internal (binary) nodes, hence n external (nullary) nodes. Let ρ be this correspondence examplified on Figure 1. The reader can convince himself easily that

$$\text{height}(t) = \text{height}^*(\rho(t)) + 1$$

where height* denotes the one-sided height of binary trees, defined as the maximum number of (internal) left branching nodes on any branch of the tree. Since for any binary tree

$$\text{height}(u) \geq \text{height}^*(u) + 1$$

it follows for the family $\mathcal{B}$ of binary trees that

$$H_{n-1}(\mathcal{B}) \geq H_n(\mathcal{G}).$$

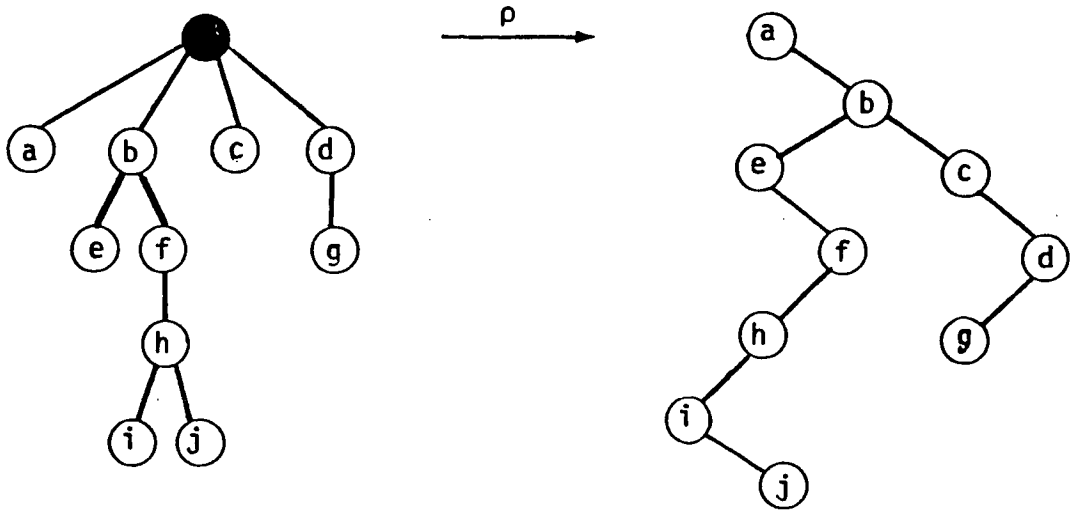

Figure 1 : The Rotation Correspondence transforms a general tree into a binary tree : the leftmost-son relation becomes the left-son relation and the right-brother relation becomes the right-son relation ; the root of the general tree is dropped. External nodes of the binary tree are not represented.

Thus the estimation of the average height of general planar trees a priori shows $H_n(B)$ to be of order $\sqrt{\pi n}$ at least.

The result about height of general trees is also of interest in another context. It is possible [8,9] to optimize the recursive visit procedure in the case of binary trees by eliminating end-recursion. The resulting iterative algorithm keeps at each stage a list of right subtrees that still remain to be explored ; the storage complexity of this optimized iterative algorithm is easily seen to correspond exactly to one-sided height. Hence, Theorem G can be expressed as :

THEOREM G' : *The iterative traversal procedure for binary trees of size* n *has average storage complexity*

$$H_{n+1}(G) \sim \sqrt{\pi n} \qquad as \quad n \to \infty .$$

Thus the expected memory complexity of the optimized iterative exploration algorithm is asymptotically (for large sizes of trees) half the expected complexity of recursive exploration.

To conclude this brief algorithmic discussion, let us mention that if the left-to-right order in the exploration need not be kept, then exploration can be reduced to a pebbling game on trees which is equivalent to register allocation. The analysis of optimal register allocation applies there, and rephrasing results of [3,5,11] one gets

THEOREM O [Optimal exploration of binary trees] : *The minimal stack size for exploring binary trees with* n *internal nodes when the left-to-right order is irrelevant, has average value*

$$\overline{O}_n = \log_4 n + P(\log_4 n) + o(1),$$

where P *is a continuous function with period* 1.

This estimation applies for instance in the context of preprocessing (allowing one bit per node).

Some comments are now in order about the relevance of our statistics : we perform analyses of tree traversal by averaging over all possible trees. The results are thus of some meaning only when inputs do not satisfy any further conditions. Basically our analyses apply to input trees with an independent labelling of nodes ; such is the case at least for expression trees in compiling, or term trees in formal manipulation systems.

As a first approximation, our treatment can also be applied to term trees in heterogeneous algebra, i-e several types of objects are present and operators have type restrictions : this corresponds to syntax trees of various sorts ; counting of trees then leads to similar statistics with generating functions that are still algebraic, and an exact treatment along our lines should be feasible (for the particular case of syntax trees of linear grammars, see [6]).

However an analysis of this type does not apply when trees occur as components of more complex structures, as appears in binary search trees or tournament trees....For instance binary search trees have monotonous labellings, and the probability distribution induced on shapes of trees by random

insertions is known [9] and far from uniform. Indeed for binary search trees, the average height for size n is O(logn) corresponding to a logarithmic search, and Robson [16] has shown the following bounds :

THEOREM BST : *Let* K_n *be the average height of binary search trees generated by* n *independent random insertions, then*

$$c_1.\log n + o(\log n) \leq K_n \leq c_2.\log n + o(\log n).$$

with $c_1>3.6$ *and* $c_2=4.31170\ldots$.

The precise asymptotic behaviour of $\frac{K_n}{\log n}$ is not yet known.

To conclude this presentation of alternative statistics, let us mention the result of Yao [17] relative to the height of index trees in dynamic hashing, which also applies to digital search trees (tries) :

THEOREM D : *Let* Γ_n *be the average height of a digital search tree constructed over* n *keys uniformly drawn on* [0,1], *then there exists constants* c_1 *and* c_2 *such that :*

$$c_1 \log n < \Gamma_n < c_2 \log n, \qquad \textit{for} \quad n\geq 2.$$

Some considerations about height in combinatorial structures are developped in our final section. We have not addressed in this paper the somewhat different problem of path length in trees, (see [8,9]) and the related question of levels of nodes in trees (which can be used to derive upper bounds on height). For this last problem the reader is referred to the excellent paper of Meir and Moon [10].

II - THE HEIGHT OF BINARY TREES : BASIC RECURRENCES

We consider the set $\mathcal{B}$ of binary trees in the sense of Knuth [8] : every node has either 0 or 2 successors and left and right successors are distinguished. The size of a tree in $\mathcal{B}$ is the number of its internal binary nodes, i-e the number of nodes with two successors ; we let $|t|$ denote the size of t. We also define

$$B_n = \text{card}\{t\in\mathcal{B} : |t|=n\}.$$

The height of a binary tree is the number of nodes along the longest branch from the root and is given inductively by

$$\begin{cases} \text{height}(\square) = 1 \\ \text{height}(t) = 1 + \max\{\text{height}(t_1), \text{height}(t_2)\} \text{ where } t_1 = \text{left}(t) \text{ and } t_2 = \text{right}(t). \end{cases}$$

Figure 2 shows the distribution of height on trees of size 4.

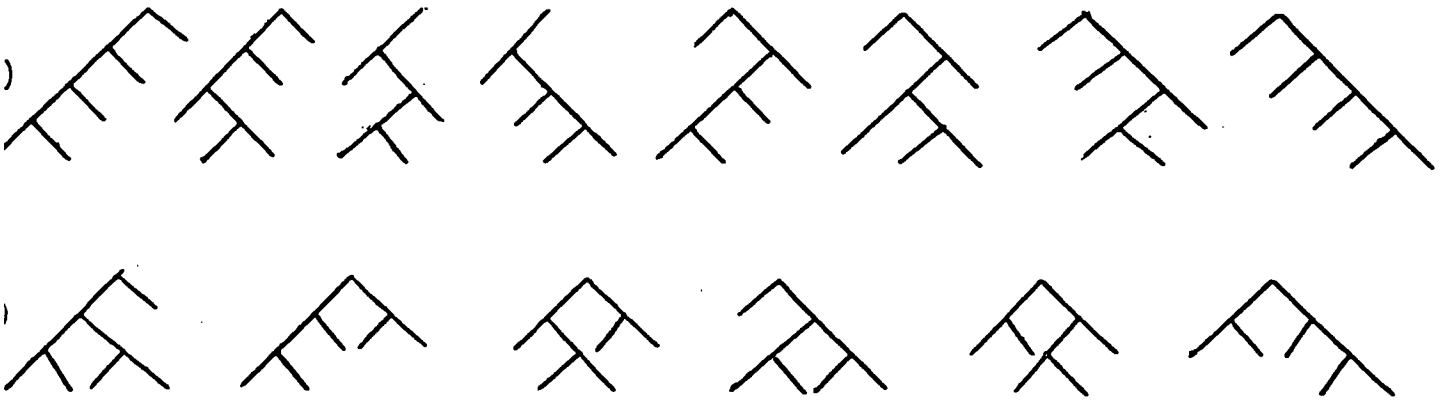

Figure 2 : Amongst the 14 trees of size 4, there are 8 trees of height 5 (a), and 6 trees of height 4 (b).

We introduce the quantities

$$B_n^{[h]} = \text{card}\{t \in \mathcal{B} : |t|=n \text{ and } \text{height}(t) \leq h\},$$

and the average height of all trees of size n, $\overline{H}_n$, is

$$\overline{H}_n = \frac{H_n}{B_n} \text{ with } H_n = \sum_{h \geq 1} h\left(B_n^{[h]} - B_n^{[h-1]}\right) . \qquad (1a)$$

From the definition, we clearly have that $B_n^{[h]} = B_n$ if $h>n$. Reorganizing the sum in (1a), we thus get

$$H_n = \sum_{h\geq 0} (B_n - B_n^{[h]}) \ . \qquad (1b)$$

The first values of these quantities are displayed in Figure 3.

n	B_n	$A_{n,1}$	$A_{n,2}$	$A_{n,3}$	$A_{n,4}$	$A_{n,5}$	$A_{n,7}$	$A_{n,7}$	H_n
1	1	1							1
2	2	0	2						2
3	5	0	1	4					2.6
4	14	0	0	6	8				3.57
5	42	0	0	6	20	16			4.24
6	132	0	0	4	40	56	32		4.87
7	429	0	0	1	68	152	144	64	5.47

Figure 3 : The distribution of height in trees of size ≤7 with $A_{n,h} = B_n^{[h]} - B_n^{[h-1]}$.

We now introduce the generating functions relative to the B_n, $B_n^{[h]}$ and H_n :

$$B(z) = \sum_{n\geq 0} B_n z^n,$$

$$B^{[h]}(z) = \sum_{n\geq 0} B_n^{[h]} z^n,$$

$$H(z) = \sum_{n\geq 0} H_n z^n.$$

The inductive definition of binary trees shows that the B_n satisfy the recurrence

$$B_n = \sum_{n_1+n_2+1=n} B_{n_1} B_{n_2}$$

whence

$$B(z) = 1 + z(B(z))^2 , \qquad (2a)$$

and

$$B(z) = \frac{1-\sqrt{1-4z}}{2z} \; ; \; B_n = \frac{1}{n+1}\binom{2n}{n} . \qquad (2b)$$

The B_n's are the Catalan numbers. From the Stirling formula follows the classical approximation :

$$B_n = \frac{4^n}{\sqrt{\pi n^3}} \left(1+0(\frac{1}{n})\right) . \qquad (2c)$$

The same decomposition principle that gives the equation for B applies to the $B^{[h]}$ showing the recurrence

$$B^{[h+1]}(z) = 1+z\left(B^{[h]}(z)\right)^2 \; ; \; B^{[0]}(z) = 0. \qquad (3)$$

No simple expression is available for the $B_n^{[h]}$ coefficients and the first values of the $B^{[h]}(z)$ are given below :

$$B^{[0]}(z) = 0 \; ; \; B^{[1]}(z) = 1 \; ; \; B^{[2]}(z) = 1+z \; ;$$

$$B^{[3]}(z) = 1+z+2z^2+z^3 \; ; \; B^{[4]}(z) = 1+z+2z^2+5z^3+6z^4+6z^5+4z^6+z^7.$$

Obviously degree $(B^{[h]}(z)) = 2^{h-1}-1$, and $B_n^{[h]} = B_n$ for $n<h$.

Summarizing the recurrences, we can state

PROPOSITION 1 : In the ring of formal power series,

$$H(z) = \sum_{h\geq 0} \left(B(z) - B^{[h]}(z)\right),$$

where B *and the* $B^{[h]}$ *satisfy :*

$$B(z) = 1+z\left(B(z)\right)^2 \; ; \; B^{[h+1]}(z) = 1+z\left(B^{[h]}(z)\right)^2 \; \textit{with} \; B^{[0]}(z) = 0.$$

We shall now proceed by proving that this expression for B(z) is also valid analytically in some domain and is a way of continuing H(z) analytically outside its circle of convergence.

PROPOSITION 2 : H(z) *has radius of convergence* $\frac{1}{4}$ *and the equality*

$$H(z) = \sum_{h\geq 0} \left(B(z) - B^{[h]}(z)\right)$$

is valid analytically inside the domain

$$C_0 = \{z : |z| \leq \tfrac{1}{4} \ \& \ z \neq \tfrac{1}{4}\} ,$$

the determination of $\sqrt{1-4z}$ *in* B(z) *being positive for real* $z < \frac{1}{4}$. *Moreover, the sum for* H(z) *converges absolutely for* z *in* C_0.

PROOF : For each non empty tree t, we have the obvious inequalities

$$1 \leq \text{height}(t) \leq |t| ,$$

which considering all trees of size n shows that

$$B_n \leq H_n \leq nB_n .$$

From the estimate (2c) of B_n follows that H(z) has radius of convergence equal to $\frac{1}{4}$.

Notice first that the series giving B(z) is absolutely convergent when $|z| \leq \frac{1}{4}$. Indeed B(z) converges as $\sum n^{-3/2}$ for all z with $|z| = \frac{1}{4}$. Let R_m denote $\sum_{n\geq m} B_n z^n$, then from simple majorations, we have

$$|B(z) - B^{[h]}(z)| \leq R_h(|z|) \text{ when } |z| \leq \tfrac{1}{4} ,$$

and $B^{[h]}(z) \to B(z)$ for any z such that $|z| \leq \frac{1}{4}$.

The nature of the convergence is obtained by writing :

$$B(z) - B^{[h+1]}(z) = z(B(z) - B^{[h]}(z))(B(z) + B^{[h]}(z)).$$

Dividing by 2 B(z) and setting $e_h(z) = \frac{B(z)-B^{[h]}(z)}{2\,B(z)}$, this recurrence is transformed into

$$e_{h+1}(z) = (1-\sqrt{1-4z})\; e_h(z)(1-e_h(z)).$$

We shall also set $\varepsilon=\varepsilon(z)=(1-4z)^{1/2}$, the determination of the square root being as above. In this notation,

$$e_{h+1}(z) = (1-\varepsilon(z))\; e_h(z)(1-e_h(z)) \quad \text{with} \quad e_0(z)=\tfrac{1}{2}\,.$$

Assuming z to be in C_0, we have $|1-\varepsilon|<1$ and the convergence of the $e_h(z)$ to 0 is geometric with

$$|e_h(z)|<c(z)\;|1-\varepsilon(z)|^h \quad \text{for some } c(z)\;;$$

thus $\sum_{h\geq 0} e_h(z)$ is also convergent and the same holds true for the sum $\sum_{h\geq 0} (B(z) - B^{[h]}(z))$. □

As will appear from considerations to be developed later, $e_n(\frac{1}{4})\sim\frac{1}{n}$ and thus $e_n(\frac{1}{4}) \to 0$ as $n \to \infty$, but at this point (where $|1-\varepsilon|=1$) the series $\sum_{n\geq 0} e_n$ diverges as the harmonic series.

In the sequel we shall mostly work with the functions $e_n(z)$. We shall thus replace equations (3), (4) by the set :

$$e_0(z)=\tfrac{1}{4} \quad e_{n+1}(z) = (1-\varepsilon(z))e_n(z)(1-e_n(z)) \tag{5}$$

and

$$H(z) = \frac{4}{1+\varepsilon(z)} \sum_{n\geq 0} e_n(z)\;; \tag{6}$$

where $\varepsilon(z) = (1-4z)^{1/2}$.

III - A FIRST ANALYTICAL CONTINUATION OF H(z) OUTSIDE THE CIRCLE OF CONVERGENCE

We proceed to show that H(z), as given by the previous recurrence equations (5), (6) is analytic in a domain larger than the circle of convergence. To that purpose, we use an argument which is essentially topological and whose principle is based on some continuity properties of a convergence criterion.

We take the complex plane cut along the axis $z>\frac{1}{4}$, $\varepsilon(z)$ being as before that branch of $\sqrt{1-4z}$ which is positive for z real, $z<\frac{1}{4}$. Consider for fixed z, the function of y :

$$f(y) = (1-\varepsilon(z))\ y(1-y),$$

in which z (or equivalently ε) enters as a parameter.

From what we have seen $e_n(z) = f^{(n)}(\frac{1}{2})$ where $f^{(n)}$ is the n-th iterate of f. We are interested in the area in which $e_n(z) \to 0$ in a non degenerate way. This can only occur if 0 is an attractive fixed point of f(y), i.e., if $f'(0) = (1-\varepsilon)$ has modulus less than 1. In this case any sequence $u_{n+1}=f(u_n)$ converges provided its initial value is close enough to the fixed point.

We thus restrict attention to values of z in the domain

$$D_0 = \{z : |1-\varepsilon(z)|<1\}.$$

Domain D_0 is the inside of a cardioid-shaped contour that properly contains C_0. The domain of values of z for which $e_n(z) \to 0$ as $n \to \infty$ thus lies somewhere between C_0 and D_0.

The following lemma is a useful convergence criterion for the sequence $\{e_m(z)\}_{m\geq 0}$.

LEMMA 1 : [Convergence criterion for $e_n(z)$]. *A necessary and sufficient condition for the sequence* $\{e_n(z)\}_{n\geq 0}$ *to converge to* 0 *for* $z\in D_0$ *is that for some* m

$$|e_m(z)| < \frac{1}{|1-\varepsilon(z)|} - 1.$$

Furthermore, if this condition is satisfied, then the convergence of the $|e_n(z)|$ *for* $n\geq m$ *is monotonic.*

PROOF : The condition of the lemma is trivially necessary. To obtain its sufficiency, note that applying the triangular inequality to the recurrence of the e_n leads to

$$|e_{n+1}| \leq |1-\varepsilon||e_n| + |1-\varepsilon||e_n|^2 ,$$

hence

$$|e_{n+1}| - |e_n| \leq |e_n||1-\varepsilon| \left(|e_n| + 1 - \frac{1}{|1-\varepsilon|}\right).$$

Thus if $|e_n| < \frac{1}{|1-\varepsilon|} - 1$, then $|e_{n+1}| < |e_n|$ and *a fortiori* $|e_{n+1}| < \frac{1}{|1-\varepsilon|} - 1$ so that the argument can be repeated. We have thus established : if for some m

$$|e_m| < \frac{1}{|1-\varepsilon|} - 1 ,$$

then for all $n \geq m$:

$$|e_{n+1}| < |e_n| < \frac{1}{|1-\varepsilon|} - 1.$$

It remains to prove that $|e_n| \to 0$ in this case. Assume *a contrario*

$$|e_n| \to L \neq 0 \quad \text{as} \quad n \to \infty.$$

Then, from the basic recurrence

$$e_{n+1} = (1-\varepsilon)\, e_n(1-e_n) ,$$

by continuity it follows that

$$|1-e_n| \to \frac{1}{|1-\varepsilon|} .$$

The conditions

$$|e_n| \to L < \frac{1}{|1-\varepsilon|} - 1 \quad \text{and} \quad |1-e_n| \to \frac{1}{|1-\varepsilon|}$$

entail that the only possible accumulation points of the sequence $\{e_n\}$ are points α satisfying

$$|\alpha| = L < \frac{1}{|1-\varepsilon|} - 1 \quad \text{and} \quad |1-\alpha| = \frac{1}{|1-\varepsilon|}$$

but these two conditions are clearly contradictory. We must therefore have L=0 which completes the proof of the proposition. □

Using the first few values of the $e_n(z)$ expressed in terms of $\varepsilon(z)$:

$$e_0(z) = \frac{1}{2} , \; e_1(z) = \frac{1}{4}(1-\varepsilon), \; e_2(z) = \frac{3}{16}(1+\frac{\varepsilon}{3})(1-\varepsilon)^2,$$

we see for instance that e_0 already satisfies the convergence criterion for $z \in [-\frac{4}{9}, \frac{2}{9}]$.

Lemma 1 can be used to show that the domain of values of z in which the sequence $\{e_n(z)\}_{n\geq 0}$ converges is an open set, and thus properly contains C_0.

LEMMA 2 : [The open set property for the convergence domain of H(z)]. *The domain* K *of values of* z *in* D_0 *for which the sequence* $\{e_n(z)\}_{n\geq 0}$ *converges is an open set.*

Furthermore the series $\sum_{n\geq 0} e_n(z)$ *is analytic in* K.

PROOF : The proof is based on the continuity of the convergence criterion of Lemma 1. If $z\epsilon K$, then for some m,

$$\phi(z) = \frac{1}{|1-\epsilon(z)|} - |e_m(z)| > 1.$$

Now clearly $\phi(z)$ is a continuous function of z inside D_0 ; thus there exists a positive real h, such that for all z' satisfying

$$|z'-z| < h,$$

we have $\phi(z')>1$. Hence $e_m(z')$ also satisfies the convergence criterion and $e_n(z') \to 0$ as $n \to \infty$.

To prove analyticity we observe that the convergence of $e_n(z)$ to 0 is geometric and uniform. Indeed since $|1-\epsilon(z)|(1+|e_m(z)|) < d < 1$ for some d, there exists a real δ such that for all z' satisfying $|z'-z| < \delta$

$$|1-\epsilon(z')|(1+|e_m(z')|) < d < 1.$$

Since for $n\geq m$ the quantities $|e_n(z')|$ decrease with n, we thus have

$$|e_n(z')| \leq d^{n-m} |e_m(z')|$$

hence $|e_n(z')| \leq c.d^n$ for some real c, uniformly in $|z'-z| < \delta$. This shows $\sum_{n\geq 0} e_n(z')$ to be uniformly convergent in $|z'-z| < \delta$, and the sum is analytic in $|z'-z| < \delta$. □

We can apply this lemma to the points in the disk $|z|\leq \frac{1}{4}$ with $z\neq \frac{1}{4}$. For each such z, there exists a $\delta(z)>0$ such that H(z) is analytic inside the domain

$$D(z) = \{z' : |z'-z| < \delta(z)\}.$$

The domain

$$D_1 = \bigcup_{z\epsilon C_0} D(z)$$

is open, properly contains C_0 and H(z) is analytic inside it.

The point $z=\frac{1}{4}$ is on the boundary of D_1, but we do not know yet the exact configuration of this boundary at $\frac{1}{4}$. However from simple topological considerations (essentially the Borel-Lebesgue lemma), we have :

PROPOSITION 3 : For each η, *there exists a* $\lambda>\frac{1}{4}$ *such that* $H(z)$ *is analytic in the indented crown*

$$|\mathrm{Arg}(z)| > \eta \quad \textit{and} \quad |z| < \lambda.$$

IV - CONTINUATION OF H(z) AROUND THE SINGULARITY

We now need to study the behaviour of the sequence $\{e_n(z)\}$ when z lies in a sector around $\frac{1}{4}$ situated inside D_0. To do so, we first show that, in part of the domain, the initial values of $e_n(z)$ decrease steadily ; we then prove that, at some stage, they satisfy the conditions of the convergence criterion (Lemma 1).

We start with the following obvious lemma.

LEMMA 3 : Let $g(y)=y(1-y)$. *If* y *satisfies*

$$|y| \le \frac{1}{4} \quad \textit{and} \quad 0 \le \mathrm{Arg}(y) \le \mathrm{Arc}\cos\frac{1}{8},$$

then $|g(y)| \le |y|$ *and* $0 \le \mathrm{Arg}\, g(y) \le \mathrm{Arg}(y)$.

PROOF : Let $y=re^{it}$. Then

$$g(y) = r(1+r^2-2r\cos t)^{1/2} \exp\left(i\left(t-\mathrm{Arc}\tan\frac{r\sin t}{1-r\cos t}\right)\right).$$

The hypothesis implies that $2r\cos t \ge r^2$ whence the bound for $|g(y)|$. On the other hand, as is easy to see,

$$0 \le \mathrm{Arc}\tan\frac{r\sin t}{1-r\cos t} \le \mathrm{Arc}\tan\sin t \le t,$$

whence the property for $\mathrm{Arg}\, g(y)$. □

LEMMA 4 : [Initial decrease of $|e_m(z)|$]. *Suppose that* $z \in D_0$, $\mathrm{Im}\, z \geq 0$, *and let* $N(z) = 1 + \left\lfloor \dfrac{\mathrm{Arc}\cos\frac{1}{8}}{\mathrm{Arg}\,(1-\varepsilon(z))} \right\rfloor$. *Then for all* $n < N(z)$,

$$|e_{n+1}(z)| \leq |e_n(z)| \leq \frac{1}{4}$$

and $0 \leq \mathrm{Arg}\,(e_{n+1}) \leq (n+1)\,\mathrm{Arg}\,(1-\varepsilon(z))$.

PROOF : It immediately follows by iterative use of the preceeding lemma. □

The restriction that $\mathrm{Im}\, z \geq 0$ in the above lemma and in the sequel is made for notational convenience since

$$e_n(\bar{z}) = \overline{e_n(z)}\ ,\ H(\bar{z}) = \overline{H(z)}\ \ldots\ .$$

We are now left with proving that for z in a certain sector around $\frac{1}{4}$, $e_{N(z)}$ satisfies the conditions of Lemma 1.

Our treatment heavily relies on a trick used by De Bruijn [1,p. 157] in the context of non-linear recurrences of a similar type. We shall express it as follows

LEMMA 5 : [Alternative recurrence on the $e_n(z)$]. *If all the* $e_j(z)$ *for* $j=0,1\ldots n-1$ *are different from* 1 *then the following relation holds :*

$$\frac{(1-\varepsilon)^n}{e_n} = \frac{1-(1-\varepsilon)^n}{\varepsilon} + 2 + \sum_{j<n} \frac{e_j}{(1-e_j)}\,(1-\varepsilon)^j. \qquad (7)$$

PROOF : We start again from the recurrence

$$e_{j+1} = (1-\varepsilon)\, e_j\, (1-e_j),$$

and we take out the $(1-\varepsilon)^j$ factor present in e_j :

$$\frac{e_{j+1}}{(1-\varepsilon)^{j+1}} = \frac{e_j}{(1-\varepsilon)^j}\,(1-e_j).$$

The essential trick now is to take inverses

$$\frac{(1-\varepsilon)^{j+1}}{e_{j+1}} = \frac{(1-\varepsilon)^j}{e_j}\,(1-e_j)^{-1}$$

and use the expansion

$$(1-u)^{-1} = 1 + u + \frac{u^2}{1-u}$$

valid provided $u \neq 1$. Here we get

$$\frac{(1-\varepsilon)^{j+1}}{e_{j+1}} = \frac{(1-\varepsilon)^j}{e_j} \left(1 + e_j + \frac{e_j^2}{1-e_j}\right),$$

$$\frac{(1-\varepsilon)^{j+1}}{e_{j+1}} = \frac{(1-\varepsilon)^j}{e_j} + (1-\varepsilon)^j + \frac{e_j}{1-e_j}(1-\varepsilon)^j.$$

When we sum these identities for $j=0..n-1$, terms like $\frac{(1-\varepsilon)^j}{e_j}$ cancel out and using the initial value $\frac{1}{e_0} = 2$, we get

$$\frac{(1-\varepsilon)^n}{e_n} = \sum_{j<n} (1-\varepsilon)^j + 2 + \sum_{j<n} \frac{e_j}{1-e_j}(1-\varepsilon)^j,$$

from which the lemma follows. □

The relation of Lemma 5 suggests $\frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n}$ as a good approximation to e_n and we are going to justify this view in the next few pages. Notice also that this relation between e_{n+1}^{-1} and e_n has the character that an upper bound on the e_j's for $j \leq n$ is turned into a lower bound on the e_{n+1}'s and *vice versa*. As an application, we study the sequence $f_n = e_n(\frac{1}{4})$ whose asymptotic behaviour will be needed later.

The f_n satisfy the recurrence

$$f_{n+1} = f_n(1-f_n) \text{ with } f_0 = \frac{1}{2},$$

hence, from Lemma 5 :

$$\frac{1}{f_n} = n + 2 + \sum_{j<n} \frac{f_j}{1-f_j}. \tag{8}$$

The f_n's being positive, it follows that

$$\frac{1}{f_n} > n+2 \quad \text{or} \quad f_n < \frac{1}{n+2}.$$

Using this more precise estimate again in (8), we get

$$\frac{1}{f_n} < n+2 + \sum_{j<n} \frac{1}{j+2} \quad .$$

Continuing the process, we see that

$$f_n = \frac{1}{n+\log n+O(1)} \quad ,$$

and more precise estimates can be derived by iteration of the process.

LEMMA 6 : [Convergence in a sector around $\frac{1}{4}$]. *There exist positive constants* ρ_0, θ_0 *such that the sequence* $\{e_n(z)\}$ *converges to* 0 *when* z *is such that*

$$z \in D_0 \; ; \; |\varepsilon(z)| < \rho_0 \quad \textit{and} \quad -\left(\frac{\pi}{4} + \theta_0\right) < \operatorname{Arg} \varepsilon(z) < -\left(\frac{\pi}{4} - \theta_0\right) \quad .$$

PROOF : We only have to show that $e_{N(z)}(z)$ is small enough to satisfy the conditions of Lemma 1. For this purpose we use Lemma 5 to provide an upper bound on $|e_{N(z)}(z)|$.

We set $\varepsilon(z)=\rho e^{i\theta}$ and expand $(1-\varepsilon(z))^{N(z)}$ in terms of ρ for small ρ when θ lies in some interval around $-\frac{\pi}{4}$ not containing 0.

The following expansions are valid for ρ small enough and $\operatorname{Arg}(\varepsilon(z)) \neq 0$. They furthermore hold uniformly when θ is in any interval of the form $[-\frac{\pi}{4} - \lambda, -\frac{\pi}{4} + \lambda]$ with $0 < \lambda < \frac{\pi}{4}$:

$$|1-\varepsilon(z)| = 1 - \rho \cos\theta + O(\rho^2),$$

$$\operatorname{Arg}(1-\varepsilon(z)) = -\rho \sin\theta + O(\rho^2),$$

$$N(z) = \frac{-\alpha}{\rho \sin\theta} + O(1) \text{ with } \alpha = \operatorname{Arc}\cos\frac{1}{8},$$

$$|1-\varepsilon(z)|^{N(z)} = e^{\alpha \cot\theta} + O(\rho).$$

In order to get an upper bound on e_N, we shall derive an asymptotic lower bound on the right hand side of the relation giving $\frac{(1-\varepsilon)^n}{e_n}$ in Lemma 5, which we take as

$$\frac{(1-\varepsilon)^n}{e_n} = \frac{1-(1-\varepsilon)^n}{\varepsilon} + \frac{8}{3} + \frac{1}{3} + \sum_{1\leq j<n} \frac{e_j}{1-e_j} (1-\varepsilon)^j.$$

Since for $1\le j\le N(z)$, $|e_j(z)| \le \frac{1}{4}$, we have $\left|\frac{e_j}{1-e_j}\right| \le \frac{1}{3}$ and :

$$\left|\frac{1}{3} + \sum_{1\le j<N} \frac{e_j}{1-e_j}(1-\varepsilon)^j\right| \le \frac{1}{3} + \frac{1}{3}\sum_{1\le j<N} |1-\varepsilon|^j$$

$$\le \frac{1}{3}\,\frac{1-|1-\varepsilon|^N}{1-|1-\varepsilon|}$$

$$< \frac{1}{3}\,\frac{1-e^{\alpha\cot\theta}}{\rho\cos\theta} + O(1)\ .$$

On the other hand

$$\left|\frac{1-(1-\varepsilon)^N}{\varepsilon}\right| \ge \frac{1-|1-\varepsilon|^N}{|\varepsilon|}$$

$$> \frac{1-e^{\alpha\cot\theta}}{\rho} + O(1).$$

Thus for ρ small enough

$$\left|\frac{1-(1-\varepsilon)^N}{\varepsilon}\right| > \frac{8}{3} + \left|\frac{1}{3} + \sum_{1\le j<N} \frac{e_j}{1-e_j}(1-\varepsilon)^j\right|\ ,$$

an inequality satisfied provided $\cos\theta > \frac{1}{3} + \delta$ for some $\delta>0$, which we shall now assume.

We have thus shown

$$\frac{|1-\varepsilon|^N}{|e_N|} > \frac{1-e^{\alpha\cot\theta}}{\rho\cos\theta}\,(\cos\theta - \tfrac{1}{3})\,(1+O(\rho))\ ,$$

or equivalently

$$|e_N| < \frac{\rho\cos\theta\,|1-\varepsilon|^N}{(1-e^{\alpha\cot\theta})(\cos\theta - \frac{1}{3})}\,(1+O(\rho)).$$

This estimate is to be compared to $\frac{1}{|1-\varepsilon|} - 1$ which is

$$\frac{1}{|1-\varepsilon|} - 1 = \rho\cos\theta + O(\rho^2)\ .$$

Thus the convergence criterion is satisfied for ρ small enough provided

$$\frac{e^{\alpha \cot \theta}}{(1-e^{\alpha i \cot \theta})} \quad \frac{1}{(\cos \theta - \frac{1}{3})} < 1 .$$

Equality is achieved for $-\theta = 0.819168... > \frac{\pi}{4}$ and inequality is ensured for all smaller values of $|\theta|$, which completes the proof of the Lemma. □

Again the convergence under the conditions of Lemma 6 is geometric except at $z = \frac{1}{4}$ and we can restate this lemma as :

PROPOSITION 4 : The function $H(z)$ *is analytic in a sector around* $\frac{1}{4}$ *defined by*

$$z \neq \frac{1}{4} \; ; \; |z-\tfrac{1}{4}| < \alpha_0 \quad \textit{and} \quad \tfrac{\pi}{2} - \beta_0 < |\mathrm{Arg}(z-\tfrac{1}{4})| < \tfrac{\pi}{2} + \beta_0,$$

for some $\alpha_0, \beta_0 > 0$.

There does not seem to be any more straightforward argument to prove convergence of $e_n(z)$ to 0 in the domain described in Propositions 3,4. Actually, numerical computations indicate that the convergence of $e_n(z)$ is not monotonic in the whole of the convergence region, and the e_n's display a fairly erratic behaviour away from the point z=1/4.

V - ESTIMATES ON H(z) AND THE AVERAGE HEIGHT OF BINARY TREES

From the results of section 3,4 as summarized by Propositions 3,4, we now know that H(z) is analytic in an indented crown shaped region depicted on Figure 4. We proceed to evaluate the Taylor coefficient H_n of H(z) by means of Cauchy's integral formula

$$H_n = \frac{1}{2i\pi} \oint H(z) \frac{dz}{z^{n+1}} ,$$

selecting a contour inside that region which gives predominance to the behaviour of the function around the singularity $\frac{1}{4}$. To do so, further information is required on the growth order of H(z) around $\frac{1}{4}$. After some preparation (Lemma 7,8), we show that H(z) behaves there like a logarithm (Proposition 5). Once this is done, we are able to conclude with the proof of Theorem B.

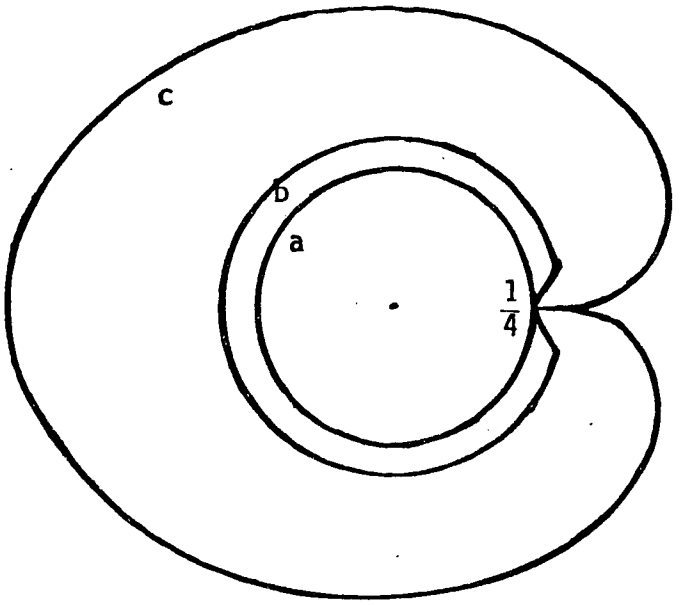

Figure 4 : A diagram representing the relative positions of the boundaries of $C_0(a)$, of $D_0(c)$ and of a convergence region guaranteed by Propositions 3,4 (b).

LEMMA 7 : [Uniform bounds for $|e_n(z)|$ around 1/4]. *There exist constants* α_1, β_1 *and* c_1 *such that*

$$|e_n(z)| < \frac{c_1}{n}$$

when $|z-\frac{1}{4}| < \alpha_1$ *and* $\frac{\pi}{2} - \beta_1 < |\mathrm{Arg}(z-\frac{1}{4})| < \frac{\pi}{2} + \beta_1$.

Moreover if† $n \geq N(z)$, *then*

$$|e_n(z)| < c_1 \, |\varepsilon(z)| \, |1-\varepsilon(z)|^n.$$

PROOF : We may suppose without loss of generality that Im $z \geq 0$. Suppose first that $1 \leq n \leq N(z)$. Let $\varepsilon(z) = \rho e^{i\theta}$. Proceeding as in the proof of Lemma 6, we find that

$$\frac{8}{3} + \left| \frac{1}{3} + \sum_{j=0}^{n-1} \frac{e_j}{1-e_j} (1-\varepsilon)^j \right| \leq \frac{8}{3} + \frac{1}{3} \, \frac{1-|1-\varepsilon|^n}{1-|1-\varepsilon|} < \frac{1}{3} \, \frac{1-|1-\varepsilon|^n}{\rho \cos\theta} + O(1),$$

---

† $N(z)$ has been defined in Lemma 4.

while

$$\left|\frac{1-(1-\varepsilon)^n}{\varepsilon}\right| \geq \frac{1-|1-\varepsilon|^n}{\rho} .$$

Hence if $n \geq c_2$, then

$$\left|3 + \sum_{j=0}^{n-1} \frac{e_j}{1-e_j} (1-\varepsilon)^j\right| < \frac{1}{2}\left|\frac{1-(1-\varepsilon)^n}{\varepsilon}\right|,$$

and so

$$\frac{|1-\varepsilon|^n}{|e_n|} \geq \frac{1}{2} \frac{|1-(1-\varepsilon)^n|}{|\varepsilon|} ,$$

$$|e_n| \leq \frac{2|\varepsilon||1-\varepsilon|^n}{1-|1-\varepsilon|^n} \leq \frac{2\rho}{1-|1-\varepsilon|^n} .$$

Take first $n \leq N(z)$; then

$$|1-\varepsilon|^n = \exp(-n\rho \cos\theta + O(n\rho^2))$$
$$\geq 1 - \delta n\rho$$

for some $\delta > 0$, and so

$$|e_n| \leq \frac{2}{\delta n} .$$

Since $|e_n| = O(n^{-1})$ for $n < c_2$, we find that

$$|e_n| \leq c_3 n^{-1} \quad \text{for} \quad n \leq N(z).$$

Let us next suppose that $n > N(z)$. Since we already know that $|e_j|$ is monotone decreasing for $j \geq N(z)$ (Lemmas 1 and 6),

$$|e_j| \leq \frac{c_3}{N(z)} \quad \text{for} \quad j \geq N(z),$$

and therefore

$$\left|3 + \sum_{j=0}^{n-1} \frac{e_j}{1-e_j} (1-\varepsilon)^j\right| \leq 3 + c_4 \sum_{j=1}^{N(z)} j^{-1} + \frac{c_3}{N(z)} \sum_{j=N(z)+1}^{n-1} |1-\varepsilon|^j$$
$$\leq c_5 \log N(z) + \frac{c_6}{N(z)} \frac{1}{\rho} \leq c_7 \log \rho^{-1} .$$

On the other hand, $|1-\varepsilon|^n \le 1/2$ for $n \ge N(z)$ and ρ small enough, so

$$\left|\frac{1-(1-\varepsilon)^n}{\varepsilon}\right| \ge \frac{1-|1-\varepsilon|^n}{\rho} \ge \frac{1}{2\rho} \quad .$$

Since $(2\rho)^{-1} > 2\, c_7 \log \rho^{-1}$ for ρ small enough,

$$|e_n| \le 4\rho\, |1-\varepsilon|^n = 4\, |\varepsilon|\, |1-\varepsilon|^n,$$

for $n \ge N(z)$ if we make α_1 small enough. This proves the last part of Lemma 7. To complete the proof of the first part, we note that for $\varepsilon = \rho e^{i\theta}$, $\frac{\pi}{2} - \beta_1 < \operatorname{Arg}\,(z - \frac{1}{4}) < \frac{\pi}{2} + \beta_1$,

$$|\varepsilon|\, |1-\varepsilon|^n \le \rho(1-\tfrac{1}{2}\rho)^n$$

and the maximum of $\rho(1-\frac{1}{2}\rho)^n$ as a function of ρ occurs at $\rho = 2(n+1)^{-1}$ and is $\le 2(n+1)^{-1}$. □

LEMMA 8 : [Uniform bound for the convergence of $e_n(z)$ to $e_n(\frac{1}{4})$]. *There exist constants* α_2, β_2 *and* c_2 *such that*

$$\left|e_n(z) - e_n(\tfrac{1}{4})\right| < c_2\, |\varepsilon(z)|$$

when $|z - \frac{1}{4}| < \alpha_2$ *and* $\frac{\pi}{2} - \beta_2 < \left|\operatorname{Arg}\,(z - \frac{1}{4})\right| < \frac{\pi}{2} + \beta_2$.

PROOF : Applying the estimate of Lemma 7 to the expansion given by Lemma 5 yields

$$\begin{aligned}\frac{(1-\varepsilon)^n}{e_n} &= \frac{1-(1-\varepsilon)^n}{\varepsilon} + O\left(\sum_{j=1}^{\infty} j^{-1}\, |1-\varepsilon|^j\right) \\ &= \frac{1-(1-\varepsilon)^n}{\varepsilon} + O(\log(1-|1-\varepsilon|)^{-1}) \\ &= \frac{1-(1-\varepsilon)^n}{\varepsilon} + O(\log|\varepsilon|^{-1}),\end{aligned}$$

as well as the already known result

$$\frac{1}{e_n(\frac{1}{4})} = n + O\left(\sum_{j<n} j^{-1}\right) = n + O(\log n).$$

Hence for $n \le N(z)$:

$$\frac{(1-\varepsilon)^n \{e_n(1/4) - e_n\}}{e_n\, e_n(1/4)} = \frac{(1-\varepsilon)^n}{e_n} - \frac{(1-\varepsilon)^n}{e_n(1/4)}$$

$$= \frac{1-(1-\varepsilon)^n - n\varepsilon(1-\varepsilon)^n}{\varepsilon} + O(\log|\varepsilon|^{-1})$$

$$= O(n^2|\varepsilon|).$$

Therefore

$$|e_n(1/4) - e_n| = O(n^2|\varepsilon\, e_n\, e_n(1/4)\, (1-\varepsilon)^{-n}|)$$

$$= O(|\varepsilon|)$$

which proves the lemma for $n \le N(z)$.

On the other hand, if $n > N(z)$, then

$$|e_n|,\ |e_n(1/4)| = O(n^{-1}) = O(|\varepsilon|),$$

so the lemma is trivial in this case. □

With these lemmas, we proceed to determine the behaviour of $H(z)$ around $\frac{1}{4}$. Our previous developments suggest approximating $\sum_{n\ge 0} e_n(z)$ by

$$L(z) = \sum_{n\ge 0} \frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n}$$

in an appropriate region. To that purpose, we study the difference

$$D(z) = \sum_{n\ge 0} e_n(z) - L(z) .$$

Using the expression for $e_n(z)$ given in Lemma 5, we see that

$$D(z) = \frac{1}{2} + \sum_{n\ge 1} e_n(z) \frac{S_n(z)}{q(n,z)} ,$$

where

$$q(n,z) = \frac{1-(1-\varepsilon(z))^n}{\varepsilon(z)} \quad \text{and} \quad S_n(z) = 3 + \sum_{1\le j<n} \frac{e_j(z)}{1-e_j(z)} (1-\varepsilon(z))^j.$$

We notice that $D(\frac{1}{4})$ exists since the defining series converges as $\sum \frac{\log n}{n^2}$. We propose to show that $D(z) = D(\frac{1}{4}) + o(1)$ as $z \to \frac{1}{4}$ and need an estimate of this $o(1)$ term.

LEMMA 9 : [First approximation lemma]. *For* z *in a neighbourhood of* $\frac{1}{4}$, *with* $|z-\frac{1}{4}| < \alpha_3$, $\frac{\pi}{2} - \beta_3 < \operatorname{Arg}(z-\frac{1}{4}) < \frac{\pi}{2} + \frac{\beta}{3}$,

$$D(z) = D(\tfrac{1}{4}) + O\left((1-4z)^{\frac{1}{4}-\eta}\right) \quad \textit{for any } \eta>0.$$

PROOF : As in Lemma 8,

$$\left|3 + \sum_{j=1}^{n-1} \frac{e_j}{1-e_j} (1-\varepsilon)^j\right| = O\left(\sum_{j=1}^{\infty} j^{-1} |1-\varepsilon|^j\right) = O(\log |\varepsilon|^{-1}).$$

However, we also have for $n\ge 3$

$$\left|3 + \sum_{j=1}^{n-1} \frac{e_j}{1-e_j} (1-\varepsilon)^j\right| = O\left(3 + \sum_{j=1}^{n-1} j^{-1}\right) = O(\log n) .$$

Therefore

$$\frac{(1-\varepsilon)^n}{e_n} = \frac{1-(1-\varepsilon)^n}{\varepsilon} + t_n ,$$

where

$$t_n = t_n(z) = O\left(\log (\min(n, |\varepsilon|^{-1}))\right).$$

Hence if n exceeds some fixed constant,

$$\frac{e_n}{(1-\varepsilon)^n} = \frac{\varepsilon}{1-(1-\varepsilon)^n} + O\left(\frac{|\varepsilon^2 t_n|}{|1-(1-\varepsilon)^n|^2}\right) ,$$

$$d_n = e_n - \frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n} = O\left(\frac{|\varepsilon^2 t_n|\ |1-\varepsilon|^n}{|1-(1-\varepsilon)^n|^2}\right) .$$

If $|\varepsilon|^{-1/2} \le n \le |\varepsilon|^{-1}$, then

$$d_n = O\left(\frac{|\varepsilon^2|\ \log n}{|1-(1-\varepsilon)^n|^2}\right) = O\left(\frac{\log n}{n^2}\right) .$$

If $n > |\varepsilon|^{-1}$,

$$d_n = O(|\varepsilon|^2\ |1-\varepsilon|^n\ \log\ |\varepsilon|^{-1}) .$$

Therefore

$$\sum_{n>|\varepsilon|^{1/2}} d_n = O\left(\sum_{n>|\varepsilon|^{-1/2}} \frac{\log n}{n^2}\right) + O\left(|\varepsilon|^2\ \log\ |\varepsilon|^{-1} \sum_{n\ge 0} |1-\varepsilon|^n\right)$$

$$= O(|\varepsilon|^{1/2}\ \log\ |\varepsilon|^{-1}) .$$

Since for all $n\ge 2$.

$$d_n(\tfrac{1}{4}) = O\left(\frac{\log n}{n^2}\right) ,$$

we find

$$\sum_{n>|\varepsilon|^{-1/2}} d_n - \sum_{n>|\varepsilon|^{-1/2}} d_n\ (\tfrac{1}{4}) = O(|\varepsilon|^{1/2}\ \log\ |\varepsilon|^{-1}).$$

For $1\le n<|\varepsilon|^{-1/2}$,

$$\frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n} = \frac{1}{n} + O(|\varepsilon|) .$$

Therefore

$$\sum_{n<|\varepsilon|^{-1/2}} (d_n - d_n(\tfrac{1}{4})) = \sum_{n<|\varepsilon|^{-1/2}} O\left(|e_n - e_n(\tfrac{1}{4})| + \left|\frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n} - \frac{1}{n}\right|\right)$$

$$= \sum_{n<|\varepsilon|^{-1/2}} O(|\varepsilon|) = O(|\varepsilon|^{1/2}) ,$$

which was to be shown. □

The constant $D(\frac{1}{4})$ in Lemma 9 can be evaluated numerically as

$$D(\tfrac{1}{4}) = \tfrac{1}{2} + \sum_{n\geq 1} (e_n(\tfrac{1}{4}) - \tfrac{1}{n}) ,$$

and we find $D_n(\frac{1}{4}) = 1.60\ldots$.

To get the final expansion of $H(z)$, we only need to estimate $L(z)$. The observation that

$$\frac{\varepsilon}{1-(1-\varepsilon)^n} \to \frac{1}{n} ,$$

for fixed n, when $\varepsilon \to 0$, suggests that $L(z)$ behaves like

$$\sum_{n\geq 1} \frac{(1-\varepsilon)^n}{n} = \log \varepsilon,$$

which we are now going to justify formally.

Notice also that expanding in powers of $(1-\varepsilon)$:

$$L(z) = \varepsilon(z) \sum_{m\geq 1} d(m)(1-\varepsilon(z))^m ,$$

with $d(m)$ the <u>divisor</u> function of m : $d(m) = \sum_{d|m} 1$.

<u>*PROPOSITION 5*</u> *:* [Main approximation lemma for $H(z)$]. *For* z *in a sector around* $\frac{1}{4}$:

$$|z - \tfrac{1}{4}| < \alpha \quad \textit{and} \quad \tfrac{\pi}{2} - \beta < \left|\mathrm{Arg}\ (z - \tfrac{1}{4})\right| < \tfrac{\pi}{2} + \beta ,$$

the following expansion holds for $H(z)$:

$$H(z) = -2 \log(1 - 4z) + K + O(|1-4z|^{\nu}) \textit{ for any } \nu < \tfrac{1}{4},$$

with $K \cong -4.1$, *a constant.*

<u>PROOF</u> : It only remains to approximate the function

$$\sum_{n\geq 1} \frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n}$$

where z is in the specified region. Setting $(1-\varepsilon) = e^{-u}$, this amounts to approximating

$$L(u) = \sum_{n\geq 1} \frac{(1-e^{-u})}{(1-e^{-nu})}\, e^{-nu} = \frac{1-e^{-u}}{u} \sum_{n\geq 1} u\, \frac{e^{-nu}}{1-e^{-nu}}$$

when u is close to 0 and Arg u is close to $\frac{\pi}{4}$.

To approximate this sum, we consider it as a Riemann sum relative to the integral

$$\int_u^\infty \frac{e^{-x}}{1-e^{-x}}\, dx .$$

Since the integral from 0 to ∞ is divergent, we split the sum according to whether $n|u| < 1$ or $n|u| \geq 1$, and compute the error terms separately.

For n such that $n|u| \geq 1$, we use the Taylor expansion

$$\left| \int_{nu}^{(n+1)u} \frac{e^{-x}}{1-e^{-x}}\, dx - \frac{ue^{-nu}}{1-e^{-nu}} \right| < \frac{|u|^2}{2} \max_{t\in[0;1]} \left| \frac{d}{dx} \frac{e^{-x}}{1-e^{-x}} \right|_{x=(n+t)u}$$

and summing, we see that

$$\sum_{n\geq |u|^{-1}} \frac{ue^{-nu}}{1-e^{-nu}} = \int_{u\lceil |u|^{-1}\rceil}^\infty \frac{e^{-x}}{1-e^{-x}}\, dx + O(|u|).$$

For n such that $n|u| < 1$, on the other hand, we expand $\frac{e^{-x}}{1-e^{-x}} - \frac{1}{x}$ which is differentiable and of bounded derivative over $[0;1]$ so that

$$\left| \frac{ue^{-nu}}{1-e^{-nu}} - \frac{1}{n} - \int_{nu}^{(n+1)u} \left(\frac{e^{-x}}{1-e^{-x}} - \frac{1}{x} \right) dx \right| < c|u|^2$$

for some constant c. Hence with $n_0 = |u|^{-1}$, we have

$$\sum_{n\geq 1} u\, \frac{ue^{-nu}}{1-e^{-nu}} = \sum_{n<\frac{1}{|u|}} \frac{1}{n} + \int_u^{n_0 u} \left(\frac{e^{-x}}{1-e^{-x}} - \frac{1}{x} \right) dx + \int_{n_0 u}^\infty \frac{e^{-x}}{1-e^{-x}}\, dx + O(|u|) .$$

Approximating the harmonic series by the logarithm and changing the bounds of the integrals with only $O(u)$ correcting terms, we see that (with γ the Euler constant) :

$$\sum_{n\geq 1} u\,\frac{e^{-nu}}{1-e^{-nu}} = -\log|u| + \gamma + \int_0^{\frac{u}{|u|}} \left(\frac{e^{-x}}{1-e^{-x}} - \frac{1}{x}\right) dx + \int_{\frac{u}{|u|}}^{\infty} \frac{e^{-x}}{1-e^{-x}}\,dx + O(|u|).$$

Using the Cauchy residue theorem, we can change the path of integration to the real axis, and we have

$$\sum_{n\geq 1}{}' u\,\frac{e^{-nu}}{1-e^{-nu}} = -\log|u| + \gamma + \int_0^1 \frac{e^{-x}}{1-e^{-x}} - \frac{1}{x}\,dx + \int_1^{\infty} \frac{e^{-x}}{1-e^{-x}}\,dx$$

$$- \int_1^{\frac{u}{|u|}} \frac{dx}{x} + O(|u|)$$

$$= -\log|u| - i\,\mathrm{Arg}\,(u) + \delta + O(|u|)$$

$$= -\log u + \delta + O(|u|)$$

with $\delta = \int_0^1 \left(\frac{e^{-x}}{1-e^{-x}} - \frac{1}{x}\right) dx + \int_1^{\infty} \frac{e^{-x}}{1-e^{-x}}\,dx + \gamma.$

In fact the two integrals cancel each other and we have $\delta = \gamma$.

Since $\varepsilon = u + U(|u|^2)$ and $\frac{1-e^{-u}}{u} = 1 + O(|u|)$, we get

$$\sum \frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n} = -\log\varepsilon + \gamma + O(|\varepsilon|)\ .$$

Combining this with the approximation in Lemma 9 yields the result, with the constant K given by

$$K = 4D_n(\tfrac{1}{4}) + 4\gamma.$$

□

To estimate the coefficients of H(z), we need to be able to translate the approximation of H(z) into an approximation of its coefficients. This is achieved through the translation lemma that follows. Since the result is of independent interest, we state it in a slightly more general form than strictly necessary here. The lemma is inspired by [12] and may be compared to the classical Darboux method although the conditions of validity differ appreciably.

PROPOSITION 6 : [Translation lemma]. *Let* G(z) *be analytic in a domain*

$$z \neq \rho \ ; \ |z| < \rho_1 \ ; \ |\mathrm{Arg}(z-\rho)| > \theta \ \textit{with} \ \ \rho_1 > \rho \quad \theta < \frac{\pi}{2} .$$

Assume G(z) *satisfies an expansion*

$$G(z) = \lambda \log \left(1-\frac{z}{\rho}\right) + \mu + \sum_{1 \le i \le m} \lambda_i \left(1-\frac{z}{\rho}\right)^{\alpha_i} + O\left(1-\frac{z}{\rho}\right)^{\nu}$$

with $0 < \alpha_1 < \alpha_2 \ldots < \alpha_m < \nu$, *valid inside the intersection of a neighbourhood of* ρ *and the domain of analycity.*

Then the n*-th Taylor coefficient* G_n *of* G(z) *admits the asymptotic expansion :*

$$G_n = \rho^{-n} \left[-\frac{\lambda}{n} + \sum_{\alpha_i + j < \nu} \frac{c_{ij}}{n^{\alpha_i + j + 1}} + O\left(\frac{1}{n^{\nu+1}}\right) \right] .$$

PROOF : The n-th Taylor coefficient can be computed using Cauchy's residue theorem, as :

$$G_n = \frac{1}{2i\pi} \int_{0^+} G(z) \ \frac{dz}{z^{n+1}}$$

where the contour simply encircles the origin and is inside the domain of analycity of the function.

We take here the specific contour

$$\Gamma(\omega) = \Gamma_{0,\omega} \cup \Gamma_{1,\omega} \cup \Gamma_2 \ ,$$

defined for some fixed θ_1 and some fixed r_1 satisfying

$$\theta < \theta_1 < \frac{\pi}{2} \quad \text{and} \quad \rho < r_1 < \rho_1 ,$$

by :

$$\Gamma_{0,\omega} = \{z : |z-\rho|=\omega \ \& \ |Arg(z-\rho)| > \theta_1$$

$$\Gamma_{1,\omega} = z : |z-\rho| \geq \omega \ \& \ |z| < \Gamma_1 \ \& \ |Arg(z-\rho)| = \theta_1\}$$

$$\Gamma_2 = \{z : |z|=r_1 \ \& \ |Arg(z-\rho)| \geq \theta_1\}.$$

The contour is depicted on Figure 5.

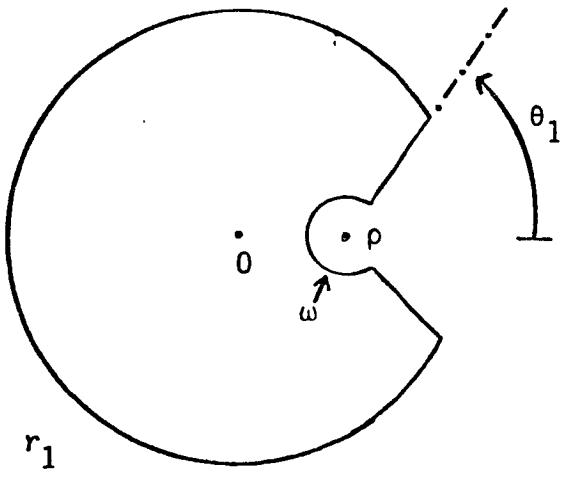

Figure 5 : A diagram showing the contour $\Gamma(\omega)$

We first shown that we can let ω shrink to zero. As ω tends to zero, the integral

$$I(\omega) = \frac{1}{2i\pi} \int_{\Gamma_{0,\omega}} G(z) \frac{dz}{z^{n+1}}$$

tends to zero, as can be seen from the inequality

$$|I(\omega)| \leq (\rho-\omega)^{-n-1} \max\{G(z) : z\in\Gamma_{0,\omega}\}.\omega \ .$$

From the local expansion follows that the upperbound vanishes with ω. Letting $\Gamma=\Gamma(0)$ and $\Gamma_1=\Gamma_{1,0}$, we thus see that G_n can be computed as :

$$G_n = \frac{1}{2i\pi} \int_{\Gamma} \frac{dz}{z^{n+1}} \ .$$

The same argument applies to the functions in the local expansion of G : $\log(1-\frac{z}{\rho})$ and the $(1-\frac{z}{\rho})^{\alpha}$, showing that

$$-\frac{\rho^{-n}}{n} = \frac{1}{2i\pi} \int_{\Gamma} \log(1-\frac{z}{\rho})\, dz$$

$$(-1)^n \rho^{-n} \binom{\alpha}{n} = \frac{1}{2i\pi} \int_{\Gamma} (1-\frac{z}{\rho})^{\alpha}\, dz.$$

Hence :

$$G_n = \rho^{-n} \left[-\frac{\lambda}{n} + \sum_{1\le i\le m} (-1)^n \binom{\alpha_i}{n} \right] + \frac{1}{2i\pi} \int_{\Gamma} R(z) \frac{dz}{z^{n+1}} \ ,$$

with

$$R(z) = G(z) - \lambda \log (1-\frac{z}{\rho}) - \mu - \sum_{1\le i\le m} \lambda_i (1-\frac{z}{\rho})^{\alpha_i} \ .$$

Now R(z) is analytic along Γ_2 and is $O(1-\frac{z}{\rho})^{\nu}$ around ρ. Consider first the integral of R(z) along Γ_2 ; we have the obvious upper bound

$$\left| \frac{1}{2i\pi} \int_{\Gamma_2} R(z) \frac{dz}{z^{n+1}} \right| < \max\{R(z) : z\in\Gamma_2\} \cdot r_1^{-n} \ .$$

R(z) being analytic along Γ_2 is bounded, and this integral is exponentially small compared to ρ^{-n}, since $r_1>\rho$. We are thus left with estimating integrals of the form

$$I_{\nu}(n) = \int_{\Gamma_2} |1-\frac{z}{\rho}|^{\nu} \frac{dz}{|z|^{n+1}} \ .$$

We set $z = \rho(1 + te^{i\phi})$ with $\phi = \pm\theta_1$ and t real ; using the symmetry of the contour, we have :

$$I_\nu(n) = 2.\rho^{-n} \int_0^\sigma \frac{t^\nu dt}{|1+te^{i\phi}|^{n+1}} \quad \text{for some } \sigma$$

$$< 2.\rho^{-n} \int_0^\infty \frac{t^\phi dt}{|1+te^{i\phi}|^{n+1}} \ .$$

Now $|1+te^{i\phi}| = (1+t^2+2t\cos\phi)^{1/2}$ and since $\cos\phi > 0$, we have $(1+t^2+2t\cos\phi)^{1/2} > 1+\lambda t$ for some $\lambda > 0$; so that

$$I_\nu(n) < 2.\rho^{-n} \int_0^\infty \frac{t^\nu dt}{(1+\lambda t)^{n+1}}$$

$$< 0\left(\rho^{-n} \int_0^\infty \frac{x^\nu dx}{(1+x)^{n+1}}\right) \ .$$

To conclude with the bound we only need to show that $\int_0^\infty \frac{x^\nu dx}{(1+x)^{n+1}}$ is $0\left(\frac{1}{n^{1+\nu}}\right)$.

Indeed

$$\int_0^\infty \frac{x^\nu dx}{(1+x)^{n+1}} = \int_0^1 \frac{x^\nu dx}{(1+x)^{n+1}} + 0(2^{-n}) \ ;$$

for $x \in [0;1]$, $(1+x) > e^{x/2}$, so that

$$\int_0^1 \frac{x^\nu dx}{(1+x)^{n+1}} < \int_0^1 x^\nu \, e^{-(\frac{n+1}{2})x} \, dx$$

$$< \int_0^1 x^\nu \, e^{-(\frac{n+1}{2})x} \, dx$$

$$< \frac{\Gamma(\nu+1)\ 2^{\nu+1}}{(n+1)^{\nu+1}} \ .$$

Hence

$$I_\nu(n) = O\left(\frac{\rho^{-n}}{n^{1+\nu}}\right) .$$

Putting everything together, we have thus shown that

$$G_n = \rho^{-n}\left[-\frac{\lambda}{n} + \sum_{1\le i\le m} \binom{\alpha_i}{n}(-1)^n\right] + O(\rho^{-n}n^{-\nu-1}).$$

To conclude the proof of the proposition, there only remains to examine the asymptotics of coefficients of the form

$$(-1)^n \binom{\alpha}{n} = (-1)^n \frac{\alpha(\alpha-1)\ldots(\alpha-n+1)}{n!}$$
$$= \frac{1}{n} \frac{\Gamma(n-\alpha)}{\Gamma(-\alpha)\,\Gamma(n)} .$$

Known properties of the gamma function show the existence of an asymptotic expansion

$$(-1)^n \binom{\alpha}{n} \sim \sum_{j\ge 0} \frac{c_j(\alpha)}{n^{\alpha+j+1}} ,$$

with, in particular,

$$c_0(\alpha) = (\Gamma(-\alpha))^{-1} .$$

Plugging these expansions into the estimate for G_n thus completes the proof of proposition 6 with

$$c_{ij} = \lambda_i\, c_j(\alpha_i). \qquad \square$$

We have thus seen that adequate local information on a function G around its singularity leads to corresponding asymptotic information on its Taylor coefficients. The better the local approximation, the more terms the asymptotic expansion contains.

We can now complete the proof of theorem B.

Proposition 3 shows H(z) to be analytic outside the circle of convergence. Proposition 4, together with proposition 5 provides the local expansion around the singularity $\frac{1}{4}$ (the expansion is actually also valid inside the circle of convergence). Hence :

THEOREM B : The average height of binary trees with n *internal nodes satisfies*

$$H_n = 2\sqrt{\pi n} + O(n^{1/4+\eta}) \quad \textit{for any} \quad \eta>0.$$

Proposition 6 also shows that any improvement in the expansion of H(z) will lead to a better error term.

n	H_n	$H_n/\sqrt{2\pi n}$
10	7.07	0.631
20	11.29	0.712
50	19.97	0.797
100	29.98	0.846
200	44.29	0.883
500	72.94	0.920
1000	105.42	0.940
2000	151.50	0.956
5000	243.17	0.970
10000	346.64	0.978
16000	440.31	0.982

Figure 6 : The average height of binary trees : Comparison of the exact values to the asymptotic estimates.

Numerical results corresponding to theorem B are displayed on Figure 6. We notice that the convergence of H_n to $\sqrt{2\pi n}$ is initially quite slow ; however for sizes of trees about 16000, the gap appears to be less than 2 %.

VI - HEIGHT IN SIMPLE FAMILIES OF TREES

Following Meir and Moon, we now consider planar trees with labels attached to nodes. All labels are taken from a fixed label set L

$$L = L_0 \cup L_1 \cup L_2 \cup \ldots ,$$

with L_r the set of labels that may be attached to a node of degree r. We assume that each of the L_r is finite and we let c_r denote $|L_r|$; we can also assume without loss of generality that all the L_r's are disjoint. A family defined in this way is said to be simple (or simply generated [10]). This definition obviously includes all families of unlabelled trees defined by restriction on the set of allowed node degrees (in which case $c_r = 0$ or 1). It also covers all families of term trees, i-e tree representations of expressions over an arbitrary set of operators. As examples, we mention

- (α) - the family of binary trees for which $c_0 = c_2 = 1$ and $c_r = 0$ for $r \neq 0;2$; these have been considered in the previous sections ;
- (β) - the family of general planar trees for which $c_r = 1$ for all $r \geq 0$: the analysis in [2] deals with these trees ;
- (γ) - the family of unary-binary trees for which $c_0 = c_1 = c_2 = 1$ and $c_r = 0$ for $r > 2$; they appear as shapes of expression trees when unary as well as binary operations are allowed ; the trees are counted by the Motzkin numbers ;
- (δ) - the family of 2-3 trees (unbalanced) for which $c_0 = c_2 = c_3 = 1$ and $c_r = 0$ otherwise ; their balanced counterparts are a useful data structure and have been counted by Odlyzko [12] ;
- (ε) - the family of t-ary trees (which also appear in digital search) ; for these trees $c_r = 1$ if $r = 0$ or t and $c_r = 0$ otherwise.

As in the above examples, we shall restrict attention to those simple families for which there exists an absolute constant M such that

$$\forall r \quad c_r < M,$$

although our treatment essentially generalises to sequences $\{c_r\}$ with a growth rate limited by an exponential.

Up to isomorphism, a simple family of trees is described by the sequence $\{c_r\}_{r\geq 0}$. Given a simple family F, we let y_n denote the number of trees of total size n - i-e the number of trees formed with a total of n nodes. The generating function

$$y(z) = \sum_{n\geq 1} y_n z^n$$

satisfies an equation of the form

$$y(z) = z\,\phi(y(z)) \quad \text{where} \quad \phi(y) = \sum_{n\geq 0} c_r y^r.$$

Also, if we define

$$y_n^{[h]} = \#\ \text{trees of size n and height} \leq h,$$

with height measured by the number of nodes along the longest branch, then the generating functions

$$y^{[h]}(z) = \sum_n y_n^{[h]}(z)$$

are given by

$$y^{[0]}(z) = 0 \qquad y^{[h+1]}(z) = z\,\phi(y^{[h]}(z)).$$

The functions ϕ corresponding to case α-ε above are thus respectively :

$$1+y^2 \ ;\ (1-y)^{-1} \ ;\ 1+y+y^2 \ ;\ 1+y^2+y^3 \ ;\ 1+y^t \ .$$

In the case of general planar trees, the $y^{[h]}(z)$ appear as convergents of a continued fraction, and additional algebraic information is available leading to explicit expressions for the $y^{[h]}(z)$; this is the basis of the treatment in [2].

In the binary case, there is a slight difference between the equation we get here, namely

$$y(z) = z(1+(y(z))^2),$$

and the equation for B(z) which is

$$B(z) = 1 + z(B(z))^2.$$

The two functions are related by

$$y(z) = z\, B(z^2),$$

which reflects the fact that in this section we consider total size measured by the total number of nodes (both nullary and binary).

The case of non planar labelled trees (with distinct labels) does not fall into our category of simple trees. It can however be subjected to the same analytical treatment since the exponential generating function

$$\hat{y}(z) = \sum y_n \frac{z^n}{n!} \quad \text{with} \quad y_n = \neq \text{ trees of size } n,$$

satisfies the equation

$$\hat{y}(z) = z \exp(\hat{y}(z)),$$

with similar expressions relative to trees of bounded height. We shall thus obtain the Renyi and Szekeres result [14] as a consequence of our theorem S.

We now indicate the lines along which the method employed for binary trees can be extended to these simple families of trees.

Let $$H_n = \sum_{h\geq 0} h(y_n^{[h]} - y_n^{[h-1]})$$

denote the total height of trees of size n, with the generating function

$$H(z) = \sum_{n\geq 0} H_n z^n.$$

We are interested in the average heights defined by

$$\overline{H}_n = \frac{H_n}{y_n},$$

provided $y_n \neq 0$. We proceed by proving that y(z) has algebraic singularities on its circle of convergence [10], and that H(z) has corresponding logarithmic singularities.

We have to distinguish two cases based on the value of

$$d = \mathrm{GCD}\ \{r \mid c_r \neq 0\}.$$

The situation where d=1 (planar trees, unary-binary ...) is the simplest one since, then, y has only one singularity on its circle of convergence ; in this case, $y_n \neq 0$ for all $n \geq n_0$. The situation where $d \neq 1$ (binary trees, t-ary trees ...) requires combining results relative to each of the d singularities of y on its circle of convergence ; in that case $y_n = 0$ if $n \not\equiv 1 \pmod d$.

Case 1 : Unicity of singularity

We start again with the equation

$$y(z) = z\ \phi(y(z))$$

and look for the point where the implicit function theorem ceases to apply. This occurs when

$$\frac{d}{dy}\left(\frac{y}{\phi(y)}\right) = 0\ ,\quad \text{i-e}\quad \phi(y) = y\,\phi'(y).$$

Let τ be the value of smallest modulus such that $\phi(\tau) = \tau\,\phi'(\tau)$. The GCD condition implies that τ is unique and real ; let $\rho = \frac{\tau}{\phi(\tau)}$ be the corresponding value of z. For (z,y) in a neighbourhood of (ρ,τ) satisfying $y = z\,\phi(y)$, a local expansion shows that

$$(z-\rho) = -(y-\tau)^2 \left(\frac{\phi''(\tau)\tau}{2\phi^2(\tau)}\right) + O((y-\tau)^3).$$

Hence around $z = \rho$, y behaves as

$$\tau - \left(\frac{2\phi(\tau)}{\phi''(\tau)}\right)^{1/2} \left(1 - \frac{z}{\rho}\right)^{1/2}$$

and its n-th Taylor coefficient is asymptotic to

$$c_1 \rho^{-n}\, n^{-3/2} \quad \text{with} \quad c_1 = \left(\frac{\phi(\tau)}{2\pi\phi''(\tau)}\right)^{1/2}.$$

This is essentially the Darboux-Polya theorem applied to tree enumerations (see [10]).

Starting from the two equations

$$y(z) = z\,\phi(y(z)) \;;\quad y^{[h+1]}(z) = z\,\phi(y^{[h]}(z)),$$

and substracting, we get

$$(y(z) - y^{[h+1]}(z)) = z(\phi(y(z)) - \phi(y^{[h]}(z)).$$

Using the Taylor expansion of the right hand side of this equality around $y(z)$, we see that

$$(y - y^{[h+1]}) = z(y - y^{[h]})\,\phi'(y)\left[1 - (y-y^{[h]})\,\frac{\phi''(y)}{2\phi'(y)} + O((y-y^{[h]})^2)\right].$$

When $z=\rho$, $z\phi'(y) = 1$; expanding $z\phi'(y)$ around ρ, we get

$$z\phi'(y) = 1 + (y-\tau)\tau\,\frac{\phi''(\tau)}{\phi(\tau)} + O(y-\tau)^2$$

$$= 1 - (1 - \frac{z}{\rho})^{1/2}\,\tau\,\left(\frac{2\phi''(\tau)}{\phi(\tau)}\right)^{1/2} + O(y-\tau)^2.$$

Thus setting $e_h(z) = y(z) - y^{[h]}(z)$, and $1 - z\phi'(y) = \varepsilon(z)$, we see that

$$e_{h+1}(z) = (1-\varepsilon(z))\;\; e_h(z)\left(1 - \frac{\phi''(\tau)}{2\phi'(\tau)}\;\; e_h(z) + O(e_h^2(z) + e_h(z)\;(y-\tau))\right),$$

where

$$\varepsilon(z) = (1 - \frac{z}{\rho})^{1/2}\,\tau\left(\frac{2\phi''(\tau)}{\phi(\tau)}\right)^{1/2} + O((y-\tau)^2).$$

The situation is then quite similar to what we had before. Taking inverses and applying De Bruijn's trick leads to the approximate expression

$$e_n(z) \sim c_2\;\; \frac{\varepsilon(z)(1-\varepsilon(z))^n}{1-(1-\varepsilon(z))^n}$$

with $c_2 = 2\frac{\phi'(\tau)}{\phi''(\tau)}$. Hence $H(z) = \sum_{n\geq 0} e_n(z)$ behaves around its singularity $z = \rho$ like $c_2 \log \varepsilon(z)$

and

$$H_n \sim \frac{1}{2} c_2 \rho^{-n} n^{-1},$$

or equivalently

$$\overline{H}_n \sim \frac{1}{2} \frac{c_2}{c_1} n^{1/2}.$$

Case 2 : Multiple singularities

We now assume that $d = GCD\{n \mid c_n \neq 0\}$ is non-trivial ($d \neq 1$). The equation

$$y = z\phi(y)$$

can then be put in the form

$$y = z\psi(y^d)$$

with $\psi(u) = \phi(u^{1/d})$ a power series in u.

The previous computations apply here : if τ is the smallest positive root of the equation

$$\phi(\tau) = \tau\,\phi'(\tau),$$

then $y(z)$ has an algebraic singularity at τ. Now, since $\phi(y)$ depends only on y^d, we see that $y(z)$ also has singularities at the points

$$\tau_j = \omega^j \tau \quad \text{for} \quad j = 0,1,\ldots,d-1,$$

where ω is a primitive d-th root of unity. Setting as before

$$\rho = \frac{\tau}{\phi(\tau)},$$

these singularities correspond to values of z

$$\rho_j = \omega^j \rho.$$

Local expansions for y can also be carried out around the ρ_j showing that

$$(z-\rho_j) = -\omega^j (y-\tau_j)^2 \frac{\phi''(\tau)\tau}{2\phi^2(\tau)} + O((y-\tau_j)^3).$$

Hence, around $z = \rho_j$, the approximation of y is

$$\tau_j - \omega^j \left(\frac{2\phi(\tau)}{\phi''(\tau)}\right) (1 - \frac{z}{\rho_j})^{1/2} .$$

The n-th Taylor coefficient of this approximation is approximated by

$$c_1 \rho^{-n} \omega^{-j(n-1)} n^{-3/2} \quad \text{with} \quad c_1 = \left(\frac{\phi(\tau)}{2\pi\phi''(\tau)}\right)^{1/2} .$$

and provided $n \equiv 1 \pmod d$ - which is to be assumed since $y_n = 0$ if $n \not\equiv 1 \pmod d$ - these terms add up to

$$d\, c_1 \rho^{-n} n^{-3/2} .$$

The same phenomenon occurs for H(z) which has also d singularities on its circle of convergence. Around $z = \rho_j$, H(z) behaves as

$$\frac{1}{2} c_2 \omega^j \log \left(1 - \frac{z}{\rho_j}\right) ,$$

so that for $n \equiv 1 \pmod d$

$$H_n \sim \frac{d}{2} c_2 \rho^{-n} n^{-1},$$

hence again

$$\mathrm{H}_n \sim \frac{1}{2} \frac{c_2}{c_1} n^{1/2} .$$

We can thus state :

<u>THEOREM S</u> : *For simple families of trees corresponding to the equation* $y = z\phi(y)$, *and for* $n \equiv 1 \pmod d$ *with* $d = \mathrm{GCD}\{r : c_r \neq 0\}$, *the average heights satisfy :*

$$\mathrm{H}_n \sim \lambda . n^{1/2} ,$$

where

$$\lambda = \left(\frac{2\pi}{\phi(\tau)\phi''(\tau)}\right)^{1/2} \phi'(\tau) ,$$

and τ *is the smallest positive root of the equation*

$$\phi(\tau) - \tau\phi'(\tau) = 0.$$

COROLLARY :

(i) *The average height of a unary-binary tree with* n *nodes is asymptotically*

$$\sqrt{3\pi n}.$$

(ii) *The average height of an unbalanced* 2-3 *tree with* n *nodes is asymptotically*

$$\sqrt{\pi \frac{2+3\tau}{1+3\tau} n}$$

where τ *is the positive root of the equation* $2\tau^3+\tau^2-1 = 0$.

(iii) *The average height of a* t*-ary tree with* n *internal (*t*-ary) nodes is asymptotically*

$$\sqrt{2\pi \frac{t}{t-1} n}$$

(iv) *The average height of a (planar rooted) tree with* n *nodes* [2] *is asymptotically*

$$\sqrt{\pi n} \quad .$$

(v) *The average height of a labelled non planar tree with* n *nodes* [14] *is asymptotically*

$$\sqrt{2\pi n} \quad .$$

VII - DISTRIBUTION RESULTS AND OTHER EXTENSIONS

In this section, we show that our methods can be extended to derive information about the distribution of height in simple families of trees. We shall deal with the binary case giving asymptotic equivalents for moments of higher order (variance...) . The distribution of height in trees appears to obey a limiting theta distribution. A similar result has been proved by Renyi and Szekeres [14] in the case of labelled non planar trees using a rather different method, and in the case of general planar trees by Kemp [7] using the explicit enumeration results available in that particular case.

We propose to prove :

THEOREM MB : [Moments of the distribution of height in binary trees] :
The r-th *moment of the distribution of height in binary trees of size* n *satisfies*

$$\overline{M}_{r,n} \sim 2^r r(r-1)\,\Gamma(\tfrac{r}{2})\zeta(r)\,n^{r/2} \qquad as \quad n \to \infty \ .$$

PROOF : The r-th moment of the distribution of height in trees of size n is given by :

$$\overline{M}_{r,n} = \frac{M_{r,n}}{B_n} \qquad \text{with} \quad M_{r,n} = \sum_{h\geq 1} h^r \left(B_n^{[h]} - B_n^{[h-1]} \right) .$$

The quantities $M_{r,n}$ are estimated from their generating function :

$$M_r(z) = \sum_{n\geq 0} M_{r,n}\, z^n ,$$

with

$$M_r(z) = \sum_{h\geq 1} h^r \left(B^{[h]}(z) - B^{[h-1]}(z) \right) .$$

We only need to consider here the case when $r>1$. Expressing M_r in terms of the e_n's and ε, we get :

$$M_r(z) = \frac{4}{1+\varepsilon(z)} \sum_{h\geq 1} h^r (e_{h-1}(z) - e_h(z))$$

$$= \frac{4}{1+\varepsilon(z)} \sum_{h\geq 0} ((h+1)^r - h^r)\, e_h(z)$$

using summation by parts. Hence setting

$$S_r(z) = \sum_{h\geq 1} h^r\, e_h(z) \ ,$$

we see that

$$M_r(z) = \frac{4}{1+\varepsilon} \left[r S_{r-1} + \binom{r}{2} S_{r-2} + \binom{r}{3} S_{r-3} \cdots \right] .$$

The problem thus reduces (for each r) to estimating the order of $M_r(z)$ around the singularity 1/4. From this information, the asymptotic behaviour of the $M_{r,n}$ is recovered by methods similar to Proposition 6.

We first compare $S_r(z)$ with the simpler function

$$T_r(z) = \sum_{n\geq 1} n^r \cdot \frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n} .$$

To do so, we study the difference $S_r - T_r$ using the tools of Lemma 9. The summation giving $S_r - T_r$ is splitted into

$$S_r - T_r = \sum_{n<|\varepsilon|^{-1/2}} n^r d_n + \sum_{|\varepsilon|^{-1/2}\leq n<|\varepsilon|^{-1}} n^r d_n + \sum_{n\geq|\varepsilon|^{-1}} n^r d_n$$

$$= U_1 + U_2 + U_3$$

with

$$d_n = e_n - \frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n} .$$

With the estimates for d_n previously derived, we find :

(i) $U_1 = O\left(\sum_{n<|\varepsilon|^{-1/2}} n^r \frac{\log n}{n^2} \right)$,

using $\left| \frac{\varepsilon(1-\varepsilon)^n}{1-(1-\varepsilon)^n} \right| = \frac{1}{n} + O(\varepsilon)$ and $t_n = O(\log \min(n, |\varepsilon|^{-1}))$.

(ii) $U_2 = O\left(\sum_{|\varepsilon|^{-1/2}\leq|\varepsilon|^{-1}} n^r \frac{\log n}{n^2} \right)$,

using $d_n = O\left(\frac{\log n}{n^2}\right)$ in this range . Hence

$U_1 + U_2 = O(\log|\varepsilon|^{-1} \cdot |\varepsilon|^{-r+1})$.

$$(iii)\quad U_3 = O\left(|\varepsilon|^2 \log |\varepsilon|^{-1} \sum_{n>|\varepsilon|^{-1}} n^r (1-\varepsilon)^n\right),$$

$$= O(\log |\varepsilon|^{-1}\ |\varepsilon|^{-r+1}),$$

$$\text{using}\quad d_n = O(|\varepsilon|^2\ |1-\varepsilon|^n \log |\varepsilon|^{-1}).$$

We have thus shown

$$|S_r - T_r| = O(\log |\varepsilon|^{-1}\ |\varepsilon|^{-r+1}),$$

a difference of a smaller order than T_r, as we now prove.

Notice first in expanding T_r that

$$T_r = \varepsilon \sum_{n\geq 1} n^r\ \frac{(1-\varepsilon)^n}{1-(1-\varepsilon)^n}$$

$$= \varepsilon \sum_{n\geq 1} \sigma_r(n)\ (1-\varepsilon)^n$$

where $\sigma_r(n)$ is the sum of the k-th powers of the divisors of n :

$$\sigma_r(n) = \sum_{d|n} d^r,$$

with corresponding Dirichlet generating function $\zeta(s)\ \zeta(s-r)$.

A function like

$$F_r(u) = \sum \sigma_r(n)\ e^{-nu}$$

can be evaluated asymptotically, for real $u \to 0$ appealing to properties of the Mellin transform as in [2]. The Mellin transform is readily found to be

$$F_r^*(s) = \zeta(s)\ \zeta(s-r)\ \Gamma(s)$$

whose rightmost pole is at $s = r+1$. From there follows by inversion, through a computation by residues

$$F_r(u) = \zeta(r+1)\ \Gamma(r+1)\ u^{-r-1} + O(u^{-1}),$$

from which $T_r(z)$ can be estimated when ε is real.

To extend this evaluation to complex z and ε, we use the method of lemma 10. We set again $e^{-u} = (1-\varepsilon)$, and

$$T_r = \varepsilon \sum_{n\geq 1} n^r \ \frac{e^{-nu}}{1-e^{-nu}}$$

$$= \varepsilon . u^{-r-1} \sum_{n\geq 1} (nu)^r \ \frac{e^{-nu}}{1-e^{-nu}} . u \quad .$$

The sum is a Riemann sum relative to the integral

$$c_r = \int_0^\infty x^r \ \frac{e^{-x}}{1-e^{-x}} \ dx \ ,$$

the integrand being of bounded derivative over the interval. We thus have :

$$T_r = c_r \varepsilon . u^{-r-1} \ (1+O(|u|)) \ ,$$

and translating back in terms of ε, we get :

$$T_r(z) = c_r \varepsilon^{-r} + O(|\varepsilon|^{-r+1}) \quad .$$

To compute c_r, it suffices to expand $(1-e^{-x})^{-1}$, and determine separately each integral. One finds :

$$c_r = \Gamma(r+1)\ \zeta(r+1) \ .$$

Returning to M_r, we have thus obtained the local expansion :

$$M_r(z) = 4r\ \zeta(r)\ \Gamma(r)\ \varepsilon^{-r+1} + O(\log|\varepsilon| . |\varepsilon|^{-r+2}) \ .$$

To conclude on the asymptotic growth of the $M_{r,n}$, we again need a translation lemma analogous to proposition 6. In fact, it is readily checked using adequate contour integration, that the bound

$$g(z) = 0\left((1 - \frac{z}{\rho})^{\alpha}\right) \qquad \alpha<0$$

implies for the n-th Taylor coefficient g_n of g the estimate†

$$g_n = O(\rho^{-n}\, n^{-\alpha-1}) \ .$$

Applying this to the error term in the expansion of $M_r(z)$, and using the explicit expressions for the coefficients of ε^{-r}, we obtain

$$M_{r,n} = 4r\zeta(r)\ \Gamma(r)\ 4^n \binom{\frac{1-r}{2}}{n} + O(4^n\, n^{r/2-5/2-\eta}) \ ,$$

for any $\eta>0$. Since for fixed non integral α

$$\binom{\alpha}{n} = \frac{1}{\Gamma(-\alpha)}\, n^{-\alpha-1}\ \left(1 + O(\frac{1}{n})\right),$$

we find :

$$M_{r,n} \sim 4r\zeta(r)\ \Gamma(r)\ \left(\Gamma(\frac{r-1}{2})\right)^{-1}\ 4^n\ n^{r/2-3/2} \ .$$

Dividing by B_n, we finally get :

$$\overline{M}_{r,n} \sim 4\pi^{1/2}\ r\ \frac{\Gamma(r)\ \zeta(r)}{\Gamma(\frac{r-1}{2})}\ \ n^{r/2} \ ,$$

which using the duplication formula for the gamma function yields

$$\overline{M}_{r,n} \sim 2^r\ r(r-1)\ \Gamma(\tfrac{r}{2})\ \zeta(r)\ n^{r/2} \ . \qquad \square$$

For n = 10.000, the asymptotic estimates of the 2nd, 3rd and 4th moment are within 10% of the actual values.

Now , we consider on binary trees, the "normalised height" defined for a tree of size n by

$$\overline{h}(t) = \frac{\text{height}(t)}{2\sqrt{n}}$$

The r-th moment $\mu_{r,n}$ of $\overline{h}$ on trees of size n satisfies

$$\mu_{r,n} \to r(r-1)\ \Gamma(\tfrac{r}{2})\ \zeta(r) \quad \text{as } n \to \infty,$$

---

† One can use a contour that circles to the left of ρ at distance $\frac{1}{n}$, then continues vertically away from ρ and closes itself at a finite distance of the circle of radius ρ.

with error terms essentially in $O(\frac{1}{n})$. (The formula is seen to be still valid for r=1). We thus see that normalized height converges to a distribution whose r-th moment is given by

$$r(r-1)\ \Gamma(\tfrac{r}{2})\ \zeta(r).$$

The limit distribution is identified by comparing these quantities with the moments of the theta distribution [14] whose cumulative distribution function is

$$H(x) = 4x^{-3}\pi^{5/2} \sum_{k\geq 0} k^2\, e^{-k^2\pi^2/x^2}$$

$$= \sum_{-\infty<k<+\infty} e^{-k^2x^2}(1-2k^2x^2)$$

with corresponding density

$$h(x) = 4x \sum_{k\geq 1} k^2(2k^2x^2-3)\, e^{-k^2x^2}\ .$$

The r-th moment of this distribution is precisely

$$\mu_r = r(r-1)\ \Gamma(\tfrac{r}{2})\ \zeta(r)\ .$$

We have thus proved :

COROLLARY : The normalized height

$$\overline{h}(t) = \frac{\text{height}(t)}{2\sqrt{n}}$$

on trees of size n *admits a limiting theta distribution with density function*

$$h(x) = 4x \sum_{k\geq 1} k^2(2k^2x^2-3)\, e^{-k^2x^2}$$

as $n \to \infty$.

The same principle applies to simple families of trees, and one finds for the r-th moment relative to trees of size n an asymptotic expression of the form

$$\xi^{r/2}\, r(r-1)\, \Gamma(\tfrac{r}{2})\, \zeta(r)\, n^{r/2}$$

which again shows that, suitably normalized, the distributions of height tend to a theta distribution.

THEOREM MS : [Moments of the distribution of height in simple trees] *For simple families of trees corresponding to the equation* $y=z\phi(y)$, *the* r*-th moment of height in trees of size* n *is asymptotic to*

$$r(r-1)\, \Gamma(\tfrac{r}{2})\, \zeta(r)\, \xi^{r/2} n^{r/2} \quad \textit{with} \quad \xi = \frac{2\phi'(\tau)^2}{\phi(\tau)\phi''(\tau)} .$$

The distribution of the normalized height in trees of size n

$$\overline{h}(t) = \frac{\text{height}(t)}{\sqrt{\xi n}}$$

tends to the limiting theta distribution of density $h(x)$.

As a matter of conclusion we would like to mention that many combinatorial problems -especially tree enumerations- have generating functions associated to functional equations of the form

$$f(z) = \Phi(z, f(z))$$

where Φ is a functional reflecting the structural definition of the objects. The approximations provided by the iterative scheme

$$f^{[0]}(z) = 0 \quad ; \quad f^{[h+1]}(z) = \Phi(z, f^{[h]}(z))$$

are often of combinatorial significance, representing a partition of the objects according to some form of "height". We have dealt previously with, equations of the form

$$f(z) = z\, \phi(f(z))$$

corresponding to simple families of trees.

The enumeration of non planar unlabelled rooted trees corresponds to equations of the form

$$f(z) = z e^{f(z) + \frac{1}{2} f(z^2) + \frac{1}{3} f(z^3) + \ldots}$$

as appears from developments in Polya theory. The present approach is applicable since the occurrence of $f(z^2)$; $f(z^3)$ ... is known not to affect singularities too much and $F(z)$ still has an algebraic singularity on its circle of convergence (see Polya [13]).

On the other hand the statistic about binary search trees and tournament trees represent equations of a different nature with probable singularities of the type of $\frac{1}{1-z} \log \frac{1}{1-z}$. We mention here the two equations

$$T(z) = 1 + \int_0^z T^2(z)\, dz$$

and

$$T(z) = \exp \int_0^z T(z)\, dz,$$

whose approximations provided by the iterative scheme are associated with respectively height and one-sided height. Methods developped here do not seem to apply to these problems.

Another line of extension of our methods is by looking at different limit distributions. In another work, the authors have shown that the limit distribution of binary trees of given height by size is Gaussian. The proof is there achieved by applying the saddle point method and investigating the analytical properties of the $B^{[h]}(z)$ outside the circle of convergence where they display a double exponential growth.

Finally we mention that the search of methods applicable, in a fairly general framework, to large classes of trees has already received some attention : Meir and Moon [10] have shown that path length in simple families of trees is essentially $\sim \alpha n\sqrt{n}$; Odlyzko [12] has dealt with functional equations of a general nature relative to balanced trees ; Flajolet, Steyaert [4] have shown that the simple backtracking algorithm for tree matching has linear average time when inputs are taken from any simple family of trees.

BIBLIOGRAPHY

[1] N. de Bruijn :
"Asymptotic Methods in Analysis", North Holland P.C., Amsterdam (1961).

[2] N. de Bruijn, D. Knuth and O. Rice :
"The Average Height of Planted Plane Trees", in Graph Theory and Computing, R-C. Read Editor, Academic Press, New York (1972) pp. 15-22.

[3] P. Flajolet, J.C. Raoult and J. Vuillemin :
"The Number of Registers Required to Evaluate Arithmetic Expressions", in Theoret. Comp. Sc. 9 (1979), pp. 99-125.

[4] P. Flajolet and J.M. Steyaert :
"On the Analysis of Tree Matching Algorithms", in 7-th ICALP Conf., Amsterdam (1980).

[5] R. Kemp :
"The Average Number of Registers Needed to Evaluate a Binary Optimally", in Acta Informatica 11, pp. 363-372 (1979).

[6] R. Kemp :
"The Average Height of a Derivation Tree Generated by a Linear Grammar in a Special Chomsky Normal Form", Saarbrucken University Report A 78/01 (1978).

[7] R. Kemp :
"On the Stack Size of Regularly Distributed Binary Trees", 6-th ICALP Conf., Udine (1979)

[8] D. Knuth :
"The Art of Computer Programming : Fundamental Algorithms", Addison Wesley, Reading (1968).

[9] D.E. Knuth :
"The Art of Computer Programming : Sorting and Searching", Addison-Wesley, Reading (1973).

[10] A. Meir and J.W. Moon :
"On the Altitude of Nodes in Random Trees", Canad. J. of Math, 30 (1978) pp. 997-1015.

[11] A. Meir, J.W. Moon and J.R. Pounder :
"On the Order of Random Channel Networks", in SIAM J. Alg. Disc. Math., 1 (1980) pp. 25-33.

[12] A. Odlyzko :
"Periodic Oscillations of Coefficients of Power Series that Satisfy Functional Equations", (to appear).

[13] G. Polya :
"Kombinatoriche Anzahlbestimmungen für Graphen, Gruppen und Chemische Verbindungen", Acta Mathematica 68 (1937) pp. 145-254.

[14] A. Renyi and G. Szekeres :
"On the Height of Trees", Australian J. of Math., 7 (1967) pp. 497-507.

[15] J. Riordan :
"The Enumeration of Trees by Height and Diameter", in IBM Journal of Research and Development, 4 (1960) pp. 473-478.

[16] J.M. Robson :
"The Height of Binary Search Trees", in The Australian Computer Journal, 11 (1979) pp. 151-153.

[17] A.C. Yao :
"A Note on the Analysis of Extendible Hashing", in Information Processing Letters, 11 (1980) pp. 84-86